

## PATENT COOPERATION TREATY

PCT

## NOTIFICATION OF ELECTION

(PCT Rule 61.2)

From the INTERNATIONAL BUREAU

To:

Assistant Commissioner for Patents  
United States Patent and Trademark  
Office  
Box PCT  
Washington, D.C.20231  
ETATS-UNIS D'AMERIQUE

in its capacity as elected Office

<b>Date of mailing (day/month/year)</b> 26 June 2000 (26.06.00)	<b>Applicant's or agent's file reference</b> P21873-PO
<b>International application No.</b> PCT/JP99/07050	<b>Priority date (day/month/year)</b> 15 December 1998 (15.12.98)
<b>International filing date (day/month/year)</b> 15 December 1999 (15.12.99)	
<b>Applicant</b> IMAGAWA, Taro et al	

1. The designated Office is hereby notified of its election made:

☒ in the demand filed with the International Preliminary Examining Authority on:

29 May 2000 (29.05.00)

☐ in a notice effecting later election filed with the International Bureau on:

2. The election ☒ was  
☐ was not

made before the expiration of 19 months from the priority date or, where Rule 32 applies, within the time limit under Rule 32.2(b).

<p>The International Bureau of WIPO 34, chemin des Colombettes 1211 Geneva 20, Switzerland</p> <p>Facsimile No.: (41-22) 740.14.35</p>	<p>Authorized officer</p> <p>R. Forax</p> <p>Telephone No.: (41-22) 338.83.38</p>
--	---

**THIS PAGE BLANK (USPTO)**

をさらに包含する、請求項 1 に記載の検索方法。

5. 前記第 1 の文字要素と前記第 2 の文字要素との間には、前記第 1 の文字要素と前記第 2 の文字要素との類似度に関連する複数の距離が予め設定されており、

5 前記複数の距離のうち選択された 1 つが前記距離として使用される、請求項 1 に記載の検索方法。

6. 前記複数の距離のうちの 1 つは、ユーザからの入力に基づいて選択される、請求項 5 に記載の検索方法。

10

7. 前記距離は、確率的な分布を有している、請求項 1 に記載の検索方法。

8. (補正後) 文字列を文字認識することによって得られる第 1 の文字要素列から第 2 の文字要素列を検索する検索方法であって、

15 前記第 1 の文字要素列は、複数の文字要素を含み、

前記複数の文字要素のうち特定の文字要素に対して、前記特定の文字要素に接続される可能性のある複数の位置の文字要素が予め決定されており、

前記検索方法は、

20 前記複数の位置の文字要素のうち特定の文字要素と、前記特定の文字要素と異なる前記複数の文字要素のうちの 1 つとを接続することによって得られる文字要素列が、前記第 2 の文字要素列の少なくとも一部に一致するか否かを判定するステップを包含する、検索方法。

9. 前記検索方法は、

25 前記特定の文字要素に接続される可能性のある前記複数の文字要素から 1 つの文字要素を選択するステップと、

**THIS PAGE BLANK (USPTO)**

前記第 1 の文字要素と前記第 2 の文字要素との間には、前記第 1 の文字要素と前記第 2 の文字要素との類似度に関連する距離が予め設定されており、

前記検索装置は、

前記距離と所定の基準距離とを比較する手段と、

- 5 前記距離と前記所定の基準距離との比較結果に基づいて、前記第 2 の文字要素が前記第 1 の文字要素に一致するか否かを判定する手段とを備えている、検索装置。

- 10 17. (補正後) 文字列を文字認識することによって得られる第 1 の文字要素列から第 2 の文字要素列を検索する検索装置であって、

前記第 1 の文字要素列は、複数の文字要素を含み、

前記複数の文字要素のうち特定の文字要素に対して、前記特定の文字要素に接続される可能性のある複数の位置の文字要素が予め決定されており、

前記検索装置は、

- 15 前記複数の位置の文字要素のうち特定の文字要素と、前記特定の文字要素と異なる前記複数の文字要素のうちの 1 つとを接続することによって得られる文字要素列が、前記第 2 の文字要素列の少なくとも一部に一致するか否かを判定する手段を備えた、検索装置。

- 20 18. 文字列を文字認識することによって得られる第 1 の文字要素列から第 2 の文字要素列を検索する検索装置であって、

前記第 1 の文字要素列は、少なくとも 1 つの第 1 の文字要素を含み、前記第 2 の文字要素列は、少なくとも 1 つの第 2 の文字要素を含み、

前記検索装置は、

- 25 前記第 2 の文字要素列に含まれる前記第 2 の文字要素の数と、前記第 2 の文字要素列に含まれる前記第 2 の文字要素のうち、前記第 1 の文字要素のうち対応す

**THIS PAGE BLANK (USPTO)**

る第1の文字要素に一致した第2の文字要素の数とに基づいて、検索結果が前記第2の文字要素列に一致する確率を取得する手段と、

前記確率に基づいて、前記検索結果の正当性を判定する手段と  
を備えた、検索装置。

5

19. 文字列を文字認識することによって得られる第1の文字要素列から第2の文字要素列を検索する検索処理をコンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、

10 前記第1の文字要素列は、第1の文字要素を含み、前記第2の文字要素列は、第2の文字要素を含み、

前記第1の文字要素と前記第2の文字要素との間には、前記第1の文字要素と前記第2の文字要素との類似度に関連する距離が予め設定されており、

前記検索処理は、

前記距離と所定の基準距離とを比較するステップと、

15 前記距離と前記所定の基準距離との比較結果に基づいて、前記第2の文字要素が前記第1の文字要素に一致するか否かを判定するステップと  
を包含する、記録媒体。

20 20. (補正後) 文字列を文字認識することによって得られる第1の文字要素列から第2の文字要素列を検索する検索処理をコンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、

前記第1の文字要素列は、複数の文字要素を含み、

前記複数の文字要素のうち特定の文字要素に対して、前記特定の文字要素に接続される可能性のある複数の位置の文字要素が予め決定されており、

25 前記検索処理は、

前記複数の位置の文字要素のうち特定の文字要素と、前記特定の文字要素と異

**THIS PAGE BLANK (USPTO)**



なる前記複数の文字要素のうちの1つとを接続することによって得られる文字要素列が、前記第2の文字要素列の少なくとも一部に一致するか否かを判定するステップ

を包含する、記録媒体。

5

21. 文字列を文字認識することによって得られる第1の文字要素列から第2の文字要素列を検索する検索処理をコンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、

前記第1の文字要素列は、少なくとも1つの第1の文字要素を含み、前記第2の文字要素列は、少なくとも1つの第2の文字要素を含み、

10

前記検索処理は、

前記第2の文字要素列に含まれる前記第2の文字要素の数と、前記第2の文字要素列に含まれる前記第2の文字要素のうち、前記第1の文字要素のうち対応する第1の文字要素に一致した第2の文字要素の数とに基づいて、検索結果が前記第2の文字要素列に一致する確率を取得するステップと、

15

前記確率に基づいて、前記検索結果の正当性を判定するステップとを包含する、記録媒体。

**THIS PAGE BLANK (USPTO)**



(51) 国際特許分類7 G06F 17/30, G06K 9/00		A1	(11) 国際公開番号 WO00/36530
			(43) 国際公開日 2000年6月22日 (22.06.00)
(21) 国際出願番号 PCT/JP99/07050		(74) 代理人 弁理士 山本秀策(YAMAMOTO, Shusaku) 〒540-6015 大阪府大阪市中央区城見一丁目2番27号 クリスタルタワー15階 Osaka, (JP)	
(22) 国際出願日 1999年12月15日 (15.12.99)			
(30) 優先権データ 特願平10/355657 1998年12月15日 (15.12.98) JP 特願平11/238053 1999年8月25日 (25.08.99) JP		(81) 指定国 CN, JP, US, 欧州特許 (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE)	
(71) 出願人 (米国を除くすべての指定国について) 松下電器産業株式会社 (MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD.) [JP/JP] 〒571-8501 大阪府門真市大字門真1006番地 Osaka, (JP)		添付公開書類 国際調査報告書	
(72) 発明者 ; および (75) 発明者 / 出願人 (米国についてのみ) 今川太郎(IMAGAWA, Taro)[JP/JP] 〒573-0071 大阪府枚方市茄子作1-9-5-102 Osaka, (JP) 松川善彦(MATSUKAWA, Yoshihiko)[JP/JP] 〒630-0213 奈良県生駒市東生駒3-207-225 Nara, (JP) 近藤堅司(KONDO, Kenji)[JP/JP] 〒573-0165 大阪府枚方市山田池東町46-3-301 Osaka, (JP) 目片強司(MEKATA, Tsuyoshi)[JP/JP] 〒576-0052 大阪府交野市私部8-18-16 Osaka, (JP)			

(54) Title: SEARCHING METHOD, SEARCHING DEVICE, AND RECORDED MEDIUM

(54) 発明の名称 検索方法、検索装置および記録媒体

	亜	啞	ㄱ	ㄴ	00
亜		10	132	166	172
啞			115	152	164
ㄱ				143	191
ㄴ					69
00					

## (57) Abstract

A method for searching for a second character element string from a first character element string obtained by recognizing a character string. The first character element string includes a first character element, and the second character element string includes a second character element. A predetermined distance between the first and second character elements is set. The distance relates to the similarity between the first and second character elements. The searching method comprises the step of comparing the distance with a predetermined reference distance and the step of judging whether or not the first character element agrees with the second element, based on the results of the comparison.

(57)要約

本発明は、文字列を文字認識することによって得られる第1の文字要素列から第2の文字要素列を検索する検索方法に関する。第1の文字要素列は、第1の文字要素を含み、第2の文字要素列は、第2の文字要素を含む。第1の文字要素と第2の文字要素との間には、第1の文字要素と第2の文字要素との類似度に関連する距離が予め設定されている。検索方法は、距離と所定の基準距離とを比較するステップと、距離と所定の基準距離との比較結果に基づいて、第1の文字要素が第2の文字要素に一致するか否かを判定するステップとを包含する。

PCTに基づいて公開される国際出願のパンフレット第一頁に掲載されたPCT加盟国を同定するために使用されるコード(参考情報)

AE	アラブ首長国連邦	DM	ドミニカ	KZ	カザフスタン	RU	ロシア
AL	アルバニア	EES	エストニア	LC	セントルシア	SD	スーダン
AM	アルメニア	EES	スペイン	LI	リヒテンシュタイン	SE	スウェーデン
AT	オーストラリア	FR	フランス	LK	スリランカ	SG	シンガポール
AU	オーストラリア	FR	フランス	LR	リベリア	SI	スロヴェニア
AZ	アゼルバイジャン	GB	英国	LS	レソト	SK	スロヴァキア
BA	ボスニア・ヘルツェゴビナ	GB	英国	LT	リトアニア	SL	シエラ・レオネ
BB	バルバドス	GD	グレナダ	LU	ルクセンブルグ	SN	セネガル
BE	ベルギー	GE	グルジア	LV	ラトヴィア	SZ	スワジランド
BF	ブルキナ・ファソ	GH	ガーナ	MA	モロッコ	TD	チャド
BG	ブルガリア	GM	ガンビア	MC	モナコ	TG	トーゴ
BJ	ベナン	GN	ギニア	MD	モルドヴァ	TJ	タジキスタン
BR	ブラジル	GW	ギニア・ビサウ	MG	マダガスカル	TZ	タンザニア
BY	ベラルーシ	GR	ギリシャ	MK	マケドニア旧ユーゴスラヴィア	TM	トルクメニスタン
CA	カナダ	HR	クロアチア		共和国	TR	トルコ
CC	中央アフリカ	HU	ハンガリー	ML	マリ	TT	トリニダード・トバゴ
CF	コンゴ	ID	インドネシア	MN	モンゴル	UA	ウクライナ
CH	スイス	IE	アイルランド	MR	モーリタニア	UG	ウガンダ
CI	コートジボワール	IL	イスラエル	MW	マラウイ	US	米国
CM	カメルーン	IN	インド	MX	メキシコ	UZ	ウズベキスタン
CN	中国	IS	アイスランド	NE	ニジェール	VN	ヴェトナム
CR	コスタ・リカ	IT	イタリア	NL	オランダ	YU	ユーゴスラビア
CU	キューバ	JP	日本	NO	ノルウェー	ZA	南アフリカ共和国
CY	キプロス	KE	ケニア	NZ	ニュージーランド	ZW	ジンバブエ
CZ	チェコ	KP	北朝鮮	PL	ポーランド		
DE	ドイツ	KR	韓国	PT	ポルトガル		
DK	デンマーク			RO	ルーマニア		

## 明 細 書

## 検索方法、検索装置および記録媒体

## 5 技術分野

本発明は、オリジナル文書を文字認識することによって得られる文書データから検索キーワードに一致する文字列を検索する検索方法、検索装置および記録媒体に関する。

## 10 背景技術

従来、文書を文字認識することによって得られる文書データから指定された文字列に基づいて関連したデータを検索する技術として、特開平7-152774号公報「文書検索方法および装置」に開示されている技術が知られている。

図23は、オリジナル文書とそれを文字認識することによって得られる文字認識結果との関係を示す。通常、文字認識を行うと、紙面に印字された文字のかすれや傾き、字体、文字サイズなどの影響により認識誤りが生じ得る。

図23に示される例では、オリジナル文書における「本」という文字が「木」という文字に誤って認識されている。また、オリジナル文書における「口」という文字が「区」という文字に誤認識されている。

20 以下、上記公報に記載の技術に基づき、図23に示される認識結果から文字列「日本」を検索する検索処理を説明する。

この検索処理には、(表1)に示されるような誤認識文字を示す表が使用される。誤認識文字を示す表は、あらかじめ、文字認識によって間違われやすい文字を並べた表である。(表1)は、「本」という文字が「木、大、太、才」に誤って認識されやすく、「口」という文字が「□(記号の四角形)、回、円、々」に誤って認識されやすいということを示している。

表 1

対象文字	誤認識文字
本	木、大、太、才
口	口、回、円、々

5

10

図 2 3 に示される認識結果から文字列「日本」を検索する場合には、誤認識文字を示す表（表 1）を用いて、文字列「日本」から文字列「日木」、「日大」、「日太」、「日才」が生成される。指定された文字列「日本」に加えて、これらの文字列「日木」、「日大」、「日太」、「日才」についても検索処理が実行される。これにより、「日本」が誤認識された「日木」の部分を検索することができるようになる。

15

しかしながら、上記公報に記載の検索処理では、誤認識しやすい文字のリストを予め用意しておくために、誤りの少ない文書データを検索する場合には、余分な文字候補を用いた余分な検索処理が行われ、逆に、誤りの多い文書データを検索する場合には、予め用意された誤認識しやすい文字のリストに含まれる文字以外の誤認識には対応できない場合が発生するという課題を有していた。

20

例えば、図 2 3 に示される例において、認識結果から文字列「人口」を検索する場合には、誤認識文字を示す表（表 1）を用いて文字列「人口（記号の四角形）」、「人回」、「人円」、「人々」が生成され、これらの文字列のそれぞれについても検索処理が実行される。しかし、誤認識文字を示す表（表 1）に存在しない誤り（「口」を「区」に間違う）が発生した場合には、「人区」（本来は「人口」）を検索することは不可能であった。

25

また、レイアウトを有する一般の文書を文字認識することによって得られる文書データから文字列を検索する場合、文字認識時のレイアウトの認識誤り（例えば、縦書き、横書きの認識誤り、改行後の次行への接続の認識誤り、段落から段

落への接続の認識誤りなど) が起こり得るが、上記公報に記載の検索処理では、レイアウトの認識誤りに対しては対応できないという課題を有していた。

例えば、図 2 4 に示されるレイアウトを有するオリジナル文書を文字認識する場合を考える。図 2 4 において段落の正しい順番は、右上の段落、左上の段落、右下の段落、左下の段落である。しかしながら、文字認識の過程において、段落の順番が誤認識されることにより、右上の段落の次に右下の段落が接続されると誤って認識される場合が起こり得る。

ここで、認識結果から文字列「日本の人口」を検索する場合には、誤認識文字を示す表などを用いて個々の文字について望ましい検索が可能である。しかし、段落の接続関係の認識が誤っている場合には、図 2 5 に示されるように「・・・日本のする傾向・・・」という認識結果として扱われるため、「日本の人口」という文字列を検索することはできない。

本発明は、上記課題に鑑みてなされたものであり、以下の (1)、(2) を目的とする。

(1) 認識結果に応じて許容可能な誤り度合を動的に変更しながら検索を行うことを可能にする検索方法、検索装置および記録媒体を提供する。

(2) 文書のレイアウトを誤って認識した場合でも、認識結果から文字列を正しく検索することを可能にする検索方法、検索装置および記録媒体を提供する。

## 発明の開示

本発明の検索方法は、文字列を文字認識することによって得られる第 1 の文字要素列から第 2 の文字要素列を検索する検索方法であって、前記第 1 の文字要素列は、第 1 の文字要素を含み、前記第 2 の文字要素列は、第 2 の文字要素を含み、前記第 1 の文字要素と前記第 2 の文字要素との間には、前記第 1 の文字要素と前記第 2 の文字要素との類似度に関連する距離が予め設定されており、前記検索方法は、前記距離と所定の基準距離とを比較するステップと、前記距離と前記所定

の基準距離との比較結果に基づいて、前記第 2 の文字要素が前記第 1 の文字要素に一致するか否かを判定するステップとを包含しており、これにより、上記目的が達成される。

5 前記第 1 の文字要素には文字認識の信頼度が予め設定されており、前記所定の基準距離は、前記信頼度に基づいて決定されてもよい。

前記所定の基準距離は、ユーザからの入力に基づいて決定されてもよい。

10 前記検索方法は、前記所定の基準距離を新たな基準距離に変更するステップと、前記距離と前記新たな基準距離とを比較するステップと、前記距離と前記新たな基準距離との比較結果に基づいて、前記第 2 の文字要素が前記第 1 の文字要素に一致するか否かを判定するステップとをさらに包含してもよい。

前記第 1 の文字要素と前記第 2 の文字要素との間には、前記第 1 の文字要素と前記第 2 の文字要素との類似度に関連する複数の距離が予め設定されており、前記複数の距離のうち選択された 1 つが前記距離として使用されてもよい。

15 前記複数の距離のうちの 1 つは、ユーザからの入力に基づいて選択されてもよい。

前記距離は、確率的な分布を有していてもよい。

20 本発明の他の検索方法は、文字列を文字認識することによって得られる第 1 の文字要素列から第 2 の文字要素列を検索する検索方法であって、前記第 1 の文字要素列は、複数の文字要素を含み、前記複数の文字要素のうち特定の文字要素に対して、前記特定の文字要素に接続される可能性のある複数の文字要素が予め決定されており、前記検索方法は、前記複数の文字要素のうち特定の文字要素と、前記特定の文字要素と異なる前記複数の文字要素のうちの 1 つとを接続することによって得られる文字要素列が、前記第 2 の文字要素列の少なくとも一部に一致するか否かを判定するステップを包含しており、これにより、上記目的が達成される。

25 前記検索方法は、前記特定の文字要素に接続される可能性のある前記複数の文



字要素から１つの文字要素を選択するステップと、前記特定の文字要素と前記選択された文字要素とを接続することによって得られる文字要素列が、前記第２の文字要素列の少なくとも一部に一致するか否かを判定するステップとを包含してもよい。

- ５ 前記特定の文字要素は、行または列の末尾に配置されており、前記特定の文字要素に接続される可能性のある前記複数の文字要素のそれぞれは、行または列の先頭に配置されていてもよい。

- １０ 前記特定の文字要素と、前記特定の文字要素に接続される可能性のある前記複数の文字要素のうちの１つとは、同一の行または列に配置されており、前記特定の文字要素と、前記特定の文字要素に接続される可能性のある前記複数の文字要素のうちの他の１つとは、異なる行または列に配置されており、かつ、同一の列または行に配置されていてもよい。

- １５ 本発明の他の検索方法は、文字列を文字認識することによって得られる第１の文字要素列から第２の文字要素列を検索する検索方法であって、前記第１の文字要素列は、少なくとも１つの第１の文字要素を含み、前記第２の文字要素列は、少なくとも１つの第２の文字要素を含み、前記検索方法は、前記第２の文字要素列に含まれる前記第２の文字要素の数と、前記第２の文字要素列に含まれる前記第２の文字要素のうち、前記第１の文字要素のうち対応する第１の文字要素に一致した第２の文字要素の数とに基づいて、検索結果が前記第２の文字要素列に一致する確率を取得するステップと、前記確率に基づいて、前記検索結果の正当性を判定するステップとを包含しており、これにより、上記目的が達成される。

- ２５ 前記第２の文字要素と前記対応する第１の文字要素との間には、前記第２の文字要素と前記対応する第１の文字要素との類似度に関連する距離が予め設定されており、前記検索方法は、前記距離と所定の基準距離とを比較するステップと、前記距離と前記所定の基準距離との比較結果に基づいて、前記第２の文字要素が前記対応する第１の文字要素に一致するか否かを判定するステップとをさらに包

含してもよい。

前記検索方法は、前記第 2 の文字要素列に含まれる前記第 2 の文字要素のうち、前記第 1 の文字要素のうち対応する第 1 の文字要素に一致しなかった第 2 の文字要素について、所定の基準距離を再設定した後に、再設定された所定の基準距離を用いて、前記第 2 の文字要素が前記対応する第 1 の文字要素に一致するか否かを再判定するステップをさらに包含してもよい。

前記検索方法は、前記第 2 の文字要素列を複数の部分に分割するステップをさらに包含してもよい。

本発明の検索装置は、文字列を文字認識することによって得られる第 1 の文字要素列から第 2 の文字要素列を検索する検索装置であって、前記第 1 の文字要素列は、第 1 の文字要素を含み、前記第 2 の文字要素列は、第 2 の文字要素を含み、前記第 1 の文字要素と前記第 2 の文字要素との間には、前記第 1 の文字要素と前記第 2 の文字要素との類似度に関連する距離が予め設定されており、前記検索装置は、前記距離と所定の基準距離とを比較する手段と、前記距離と前記所定の基準距離との比較結果に基づいて、前記第 2 の文字要素が前記第 1 の文字要素に一致するか否かを判定する手段とを備えており、これにより、上記目的が達成される。

本発明の他の検索装置は、文字列を文字認識することによって得られる第 1 の文字要素列から第 2 の文字要素列を検索する検索装置であって、前記第 1 の文字要素列は、複数の文字要素を含み、前記複数の文字要素のうち特定の文字要素に対して、前記特定の文字要素に接続される可能性のある複数の文字要素が予め決定されており、前記検索装置は、前記複数の文字要素のうち特定の文字要素と、前記特定の文字要素と異なる前記複数の文字要素のうちの 1 つとを接続することによって得られる文字要素列が、前記第 2 の文字要素列の少なくとも一部に一致するか否かを判定する手段を備えており、これにより、上記目的が達成される。

本発明の他の検索装置は、文字列を文字認識することによって得られる第 1 の

文字要素列から第2の文字要素列を検索する検索装置であって、前記第1の文字要素列は、少なくとも1つの第1の文字要素を含み、前記第2の文字要素列は、少なくとも1つの第2の文字要素を含み、前記検索装置は、前記第2の文字要素列に含まれる前記第2の文字要素の数と、前記第2の文字要素列に含まれる前記第2の文字要素のうち、前記第1の文字要素のうち対応する第1の文字要素に一致した第2の文字要素の数とに基づいて、検索結果が前記第2の文字要素列に一致する確率を取得する手段と、前記確率に基づいて、前記検索結果の正当性を判定する手段とを備えており、これにより、上記目的が達成される。

本発明の記録媒体は、文字列を文字認識することによって得られる第1の文字要素列から第2の文字要素列を検索する検索処理をコンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、前記第1の文字要素列は、第1の文字要素を含み、前記第2の文字要素列は、第2の文字要素を含み、前記第1の文字要素と前記第2の文字要素との間には、前記第1の文字要素と前記第2の文字要素との類似度に関連する距離が予め設定されており、前記検索処理は、前記距離と所定の基準距離とを比較するステップと、前記距離と前記所定の基準距離との比較結果に基づいて、前記第2の文字要素が前記第1の文字要素に一致するか否かを判定するステップとを包含しており、これにより、上記目的が達成される。

本発明の他の記録媒体は、文字列を文字認識することによって得られる第1の文字要素列から第2の文字要素列を検索する検索処理をコンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、前記第1の文字要素列は、複数の文字要素を含み、前記複数の文字要素のうち特定の文字要素に対して、前記特定の文字要素に接続される可能性のある複数の文字要素が予め決定されており、前記検索処理は、前記複数の文字要素のうち特定の文字要素と、前記特定の文字要素と異なる前記複数の文字要素のうちの1つとを接続することによって得られる文字要素列が、前記第2の文字要素列の少なく

とも一部に一致するか否かを判定するステップを包含しており、これにより、上記目的が達成される。

本発明の他の記録媒体は、文字列を文字認識することによって得られる第1の文字要素列から第2の文字要素列を検索する検索処理をコンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、  
5 前記第1の文字要素列は、少なくとも1つの第1の文字要素を含み、前記第2の文字要素列は、少なくとも1つの第2の文字要素を含み、前記検索処理は、前記第2の文字要素列に含まれる前記第2の文字要素の数と、前記第2の文字要素列に含まれる前記第2の文字要素のうち、前記第1の文字要素のうち対応する第1  
10 の文字要素に一致した第2の文字要素の数とに基づいて、検索結果が前記第2の文字要素列に一致する確率を取得するステップと、前記確率に基づいて、前記検索結果の正当性を判定するステップとを包含しており、これにより、上記目的が達成される。

#### 15 図面の簡単な説明

図1は、本発明の実施の形態1～実施の形態14に共通の検索装置1の構成を示すブロック図である。

図2は、文字要素間の距離テーブルの一例を示す図である。

図3Aは、1つの文字片からなる文字要素の例を示す図である。

20 図3Bは、1つの文字および1つの文字片からなる文字要素の例を示す図である。

図4は、オリジナル文書とそのオリジナル文書を文字認識することによって得られる認識結果との関係を示す図である。

図5は、文書データ（認識結果）に含まれる各文字要素に対して文字認識の信頼度が検索パラメータとして予め設定されている例を示す図である。  
25

図6は、「人」・「口」・「構」・「成」の各文字要素と他の文字要素との距

離関係の一部を示す図である。

図 7 は、「明朝体」というフォント種に対応するように用意された距離テーブルの例を示す図である。

5 図 8 A は、複数の文字要素を単一の文字要素として扱った結果生じる誤認識の例を示す図である。

図 8 B は、単一の文字要素を複数の文字要素として扱った結果生じる誤認識の例を示す図である。

図 9 は、図 8 A および図 8 B に示される文字要素を含む文字要素間の距離テーブルの一例を示す図である。

10 図 10 A は、発生頻度付きの距離テーブルの例を示す図である。

図 10 B は、発生頻度付きの距離テーブルの他の例を示す図である。

図 10 C は、発生頻度付きの距離テーブルの他の例を示す図である。

図 11 は、文字認識の信頼度、基準距離および発生頻度の関係の一例を示す図である。

15 図 12 は、文字認識の信頼度、基準距離および発生頻度の関係の一例を示す図である。

図 13 は、発生頻度付きの距離テーブルを作成する処理の手順を示すフローチャートである。

20 図 14 は、発生頻度付きの距離テーブルを用いて、認識結果から検索キーワードの文字要素列を検索する検索処理の手順を示すフローチャートである。

図 15 は、認識結果の候補が追加された例を示す図である。

図 16 は、段落 A ～段落 D を有するオリジナル文書の例を示す図である。

図 17 は、段落間の接続関係の認識結果を示すテーブルの例を示す図である。

25 図 18 は、段落間の接続関係を考慮して、認識結果から検索キーワードの文字要素列を検索する検索処理の手順を示すフローチャートである。

図 19 は、段落間の接続関係の認識結果を示すテーブルの例を示す図である。

図 2 0 は、オリジナル文書を文字認識することによって得られる認識結果の例を示す図である。

図 2 1 は、横書きのレイアウトを有するオリジナル文書の例を示す図である。

5 図 2 2 A は、オリジナル文書の文章が縦書きであるという仮定の下で文字認識を実行した場合の認識結果を示すテーブルの例を示す図である。

図 2 2 B は、オリジナル文書の文章が横書きであるという仮定の下で文字認識を実行した場合の認識結果を示すテーブルの例を示す図である。

図 2 3 は、オリジナル文書とそれを文字認識することによって得られる文字認識結果との関係を示す図である。

10 図 2 4 は、レイアウトを有するオリジナル文書の例を示す図である。

図 2 5 は、オリジナル文書とそれを文字認識することによって得られる文字認識結果との関係を示す図である。

図 2 6 は、認識結果から検索語を検索する検索処理を説明するための図である。

図 2 7 は、確率テーブルの一例を示す図である。

15 図 2 8 は、認識結果から検索語を検索する検索処理を説明するための図である。

図 2 9 は、認識結果から検索語を検索する検索処理を説明するための図である。

図 3 0 は、あいまい検索処理の手順を示すフローチャートである。

図 3 1 は、認識結果から検索語を検索する検索処理を説明するための図である。

## 20 発明を実施するための最良の形態

以下、図面を参照しながら、本発明の実施の形態を説明する。

図 1 は、本発明の実施の形態 1 ～実施の形態 1 4 に共通の検索装置 1 の構成を示す。

25 検索装置 1 は、端末 1 0 0 と、登録処理および検索処理を実行する CPU 1 1 0 と、文書を画像データとして入力する画像入力機器 1 2 0 と、その画像データを格納するメモリ 1 3 0 と、その画像データを文字認識することによって得られ

る文書データ（認識結果）を格納するメモリ 140 と、文字要素間の距離テーブルを格納するメモリ 150 と、文字認識用のパターン辞書 160 と、文書登録プログラム、文字認識プログラム、文書検索プログラムを格納するメモリ 170 と、ワークメモリ 180 とを含む。

- 5        検索装置 1 の各構成要素は、内部バスを介して相互に接続されてもよいし、ネットワークを介して相互に接続されてもよい。

はじめに、登録処理の流れを説明する。

- ユーザが端末 100 から登録処理の開始を指示すると、メモリ 170 上の文書登録プログラムが起動され、ワークメモリ 180 にロードされる。CPU 110  
10        は、文書登録プログラムを実行する。その結果、画像入力機器 120 によって文書が画像データとして読み取られ、その画像データがメモリ 130 に格納される。画像入力機器 120 は、例えば、スキャナ、デジタルカメラ、ビデオカメラである。

- 文書登録プログラムからメモリ 170 上の文字認識プログラムが起動され、ワークメモリ 180 にロードされる。CPU 110 は、文字認識プログラムを実行  
15        する。その結果、メモリ 130 に格納された画像データが読み出され、その画像データに含まれる文字情報を文字コード列に変換することによって文書データ（認識結果）が得られる。文書データ（認識結果）は、メモリ 140 に格納される。画像データに含まれる文字情報を文字コード列に変換する際に文字認識用の  
20        パターン辞書 160 が参照される。

- 文字認識プログラムのアルゴリズムとしては、任意のアルゴリズムが採用され得る。例えば、文字認識プログラムのアルゴリズムとして、1 文字単位に画像データを切り出し、その切り出された 1 文字単位の画像データを文字コードに変換していくというアルゴリズムを採用してもよい。

- 25        次に、検索処理の流れを説明する。

ユーザが端末 100 から検索キーワードを入力し、検索処理の開始を指示する

と、メモリ 170 上の文書検索プログラムが起動され、ワークメモリ 180 にロードされる。CPU 110 は、文書検索プログラムを実行する。その結果、メモリ 150 に格納されている距離テーブルを用いて、メモリ 140 に格納されている文書データ（認識結果）中に検索キーワードに対応する文字要素列が存在するか否かが判定される。その検索結果は、端末 100 に表示される。その検索結果とともに、その検索結果に対応する画像データを端末 100 に表示するようにしてもよい。

（実施の形態 1）

図 2 は、メモリ 150 に格納される文字要素間の距離テーブルの一例を示す。文字要素間の距離テーブルは、文字要素間の近い・遠いの関係を数値で表現したものである。

ここで、本明細書において「文字要素」とは、1 以上の文字、または、1 以上の文字片、または、1 以上の文字および 1 以上の文字片の組み合わせをいうと定義する。

例えば、「亜」は 1 つの文字からなる文字要素の例であり、「00」は 2 つの文字からなる文字要素の例である。

文字片とは、文字の一部をいう。例えば、漢字の「へん」や「つくり」が文字片に相当する。例えば、図 3 A は、1 つの文字片からなる文字要素の例を示し、図 3 B は、1 つの文字および 1 つの文字片からなる文字要素の例を示す。

なお、文字には「）」や「◎」のような記号も含まれるものとする。

図 2 に示される距離テーブルには、文字要素と他の文字要素との間の距離が予め設定されている。この距離は、文字要素間の類似度に関連する。距離が大きいほど類似度は小さく、距離が小さいほど類似度は大きい。従って、文字要素間の類似度と文字要素間の距離とは逆数の関係にある。

図 2 に示される例では、文字要素「亜」と文字要素「唾」との距離は 10 に設定されており、文字要素「亜」と文字要素「00」との距離は 172 に設定され



ている。このことは、文字要素「亜」が文字要素「0 0」よりも文字要素「哑」に類似していることを示している。他の文字要素間にも、同様にして距離が予め設定されている。

5 距離としては、文字要素間の類似度に関連する任意の値を使用することが可能である。例えば、距離として、特定の文字認識システムの入出力関係、各文字要素の形状を特徴量数値で表現した場合の特徴量空間内でのユークリッド距離などを用いることができる。

10 なお、文字要素間の距離は、必ずしも、図2に示されるような格子状の表の形式で表現されている必要はない。文字要素間の距離は、その距離が文字要素間の類似度に関連するように予め設定されている限り、任意の形式で表現され得る。例えば、距離テーブルは、文字要素ごとに他の文字要素を距離の近い順番に保持するようにしてもよく、その順番そのものを距離として扱うようにしてもよい。

以下、文字要素間の距離テーブルを用いた検索方法を説明する。

15 図4は、オリジナル文書とそのオリジナル文書を文字認識することによって得られる認識結果との関係を示す。図4に示される例では、「・・・日本の人口構成は・・・」というオリジナル文書を文字認識することによって「・・・日本の人区構成は・・・」という認識結果が得られている。この認識結果は、文書データとしてメモリ140に格納される。メモリ140は、任意のタイプの記憶媒体であり得る。

20 通常、文字認識技術を用いて文字認識を行う場合には、様々な要因で文字認識の誤りが生じ得る。図4に示される例では、「本」という文字が誤って「木」という文字に認識されており、「口」という文字が誤って「区」という文字に認識されている。

25 ここで、「日本」という文字要素列を検索キーワードとして指定して、図4に示される認識結果の中から検索キーワードに対応する文字要素列を検索する検索処理を説明する。この検索処理は、CPU110によって文書検索プログラムに

従って実行される。

まず、指定された文字要素列「日本」のうち文字要素「日」について文字要素間の距離テーブルが参照される。次に、文字要素「日」との距離が所定の基準距離（例えば、150）よりも小さい文字要素（例えば、文字要素「日」について  
5 は文字要素「日」と文字要素「目」）が図4に示される認識結果の中から検索される。この場合、認識結果の中の「日」という文字要素が検索結果として検出される。

次に、指定された文字要素列「日本」のうち次の文字要素「本」について文字要素間の距離テーブルが参照される。次に、文字要素「本」との距離が所定の基準距離（例えば、150）よりも小さい文字要素（例えば、文字要素「本」につ  
10 いては文字要素「本」と文字要素「木」と文字要素「大」）が文字要素「日」が検出された位置の次の位置の文字要素「木」に一致するか否かが判定される。この場合、文字要素「木」が、文字要素「日」が検出された位置の次の位置の文字要素に一致することから、指定された文字要素列「日本」に対して図4に示される  
15 認識結果中の文字要素列「日木」を検出することができる。

このように、本来「日本」という文字列が誤って「日木」と認識された場合でも、本来の文字列の位置を検索することが可能となる。

実際には、指定された文字要素列が検出された場合には、検出された文字要素列のみならず、検出された文字要素列を含む前後の文書の認識結果も合わせて検索者に提示したり、文字認識を行ったオリジナルの文書を画像イメージとして別途文書データに保持しておき、対応する画像イメージを検索者に提示することが  
20 好ましい。これにより、文字認識の結果が一部誤っていた場合でも、人間が再度判断することで、検索者が必要とする情報を得ることが可能となる。

また、指定された文字要素列が検出された場合には、前後の文書の提示以外に  
25 文書のタイトルや要約を表示してもよい。この場合、少ない表示スペースで検索結果を把握することが可能となる。また、表示以外に音声を用いて前後の文章や

タイトル、要約を出力することで、表示領域の少ない端末にも対応することができる。また、検索結果は通信路（ネットワーク）を経由して出力してもよい。また、帯域が狭い通信路を経由する場合には、検索結果の画像イメージを最初から表示するのではなく、前後の文書の認識結果やタイトル・要約のみを最初に表示し、検索者が別途指示することで情報量の多い画像イメージの表示を行うことで、  
5 検索時間や閲覧時間を節約することが可能となる。

更に、指定された文字要素列が検出された場合には、検出された情報を提示するのではなく、機器に新たな命令（コマンド）を発行するようにしてもよい。例えば、カメラなどからリアルタイムに得られる画像に対して検索を行い、特定の  
10 文字要素列（例えば「レストラン」）を検出した場合に撮像を行う機器に対して映像をメモリに記録するコマンドを発行するようにしてもよい。これにより、レストランの映像を自動的に集めることが可能となる。

特定の文字要素列が検索された場合には、その特定の文字要素列を含む画像をプリンタに印刷するコマンドをそのプリンタに発行したり、その特定の文字要素  
15 列を含む画像の情報を通信網（ネットワーク）を通して複数の宛名に配信したりしてもよい。

なお、文字要素間の距離と比較される基準距離は150という値に限定されない。基準距離は、任意の値に設定され得る。また、基準距離は、必ずしも固定である必要はなく、可変であってもよい。基準距離は、ユーザからの入力に基づいて  
20 決定されてもよく、CPU110によって実行される演算結果に基づいて決定されてもよい。

例えば、基準距離を最初に小さい値に設定して検索を行った結果、認識結果の中から検索キーワードに対応する文字要素列を検出することができない場合には、基準距離を順次大きな値に再設定して検索を行うようにしてもよい。これは、最初  
25 は文字認識の誤りをあまり許容しない状態で検索を行い、順次文字認識の誤りを許容して検索することに相当する。これにより、最初から文字認識の誤りを大

きく許容することによって、検索キーワードに関係の無い余分な文字要素列が検出されることを未然に防ぐことが可能となる。

また、オリジナル文書を文字認識することによって得られる文書データ（認識結果）とともに、文字認識の信頼度（又は尤度、確度、確からしさなど）を検索パラメータとして保持しておくことにより、検索パラメータに応じて検索に用いる距離の基準値（基準距離）を適切な値に設定するようにしてもよい。検索パラメータは、例えば、メモリ 140 に格納される。

図 5 は、文書データ（認識結果）に含まれる各文字要素に対して文字認識の信頼度が検索パラメータとして予め設定されている例を示す。ここで、信頼度は、0 から 1 までの値によって表される。信頼度の値が大きいほど認識結果が確からしいことを示す。

以下、「人口構成」という文字要素列を検索キーワードとして指定して、図 5 に示される認識結果の中から検索キーワードに対応する文字要素列を検索する検索処理を説明する。この検索処理は、CPU 110 によって文書検索プログラムに従って実行される。

図 6 は、「人」・「口」・「構」・「成」の各文字要素と他の文字要素との距離関係の一部を示す。

図 5 に示されるように、認識結果中の文字要素「人」・「成」には信頼度 0.9 が予め設定されている。このように高い信頼度が予め設定されている文字要素に対しては、文字要素間の距離と比較される基準距離は低く設定される。例えば、図 5 に示される例では、信頼度 0.9 の文字要素に対しては、基準距離は 10 に設定されている。認識結果中の文字要素「区」には信頼度 0.4 が予め設定されている。このように低い信頼度が予め設定されている文字要素に対しては、文字要素間の距離と比較される基準距離は高く設定される。例えば、図 5 に示される例では、信頼度 0.4 の文字要素に対しては、基準距離は 60 に設定されている。

このように、文字認識の信頼度に基づいて基準距離を可変に設定することによ

り、検索の精度を向上させることができる。例えば、認識結果の中で誤認識されている文字要素「区」の信頼度は低いため、基準距離が高く設定される。これにより、文字要素「区」との距離が50である文字要素「口」が検索対象となる。その結果、「人口構成」という文字要素列を指定して、誤認識文字を含む「人区構成」という文字要素列を検索することが可能となる。

このように、文字認識の信頼度の低い文字要素または文書については、文字要素間の距離テーブルにおいて検索時に許容される文字要素間の距離（基準距離）が大き目に設定され、逆に文字認識の信頼度が高い文字要素または文書については、文字要素間の距離テーブルにおいて検索時に許容される文字要素間の距離（基準距離）が小さ目に設定される。これにより、検索キーワードに関係の無い余分な文字列要素の検出を抑えることが可能となる。

なお、信頼度の値と文字要素間の距離テーブルにおいて許容される距離（基準距離）との対応関係は予め設定される。

さらに、文字認識の信頼度が特に低い場合には、すべての文字要素の可能性を考慮するように検索手法を切り替えるようにしてもよい。

検索パラメータ（信頼度）は、文書毎や文字要素毎に付与されてもよい。また、文字認識の信頼度としては、文字認識を行うシステム（例えば、ニューラルネットワーク）の出力や認識候補の数などを用いることが可能である。

ここでは、検索キーワードとして指定された文字要素列を構成する文字要素の先頭「日」から順番に1文字ずつ検索する例を説明したが、これと異なる順番で1文字ずつ検索するようにしてもよい。特に一般的な文書中に出現する頻度を考慮して、検索キーワードとして指定された文字要素列を構成する文字要素のうち、一般的に文書中に出現する頻度の低い文字要素から検索することにより、余分な検索手続きを減らすことが可能となり、検索速度を速めることが可能となる。

なお、上述した例では、認識結果の文書データを記憶媒体（メモリや磁気ディスク、光ディスクなど）に予め格納しておくことを想定したが、画像入力機器

(スキャナ、デジタルカメラ、ビデオカメラなど) から入力した画像データを逐次文字認識して得られるリアルタイムの情報について同様の検索を行ってもよい。

このように、文字要素間の距離テーブルを用いて文字要素列の検索を行うことにより、指定された文字要素列が誤認識で他の文字要素列に置き換わっている場合でも、指定された文字要素列に対応する文字要素列を文書データの中から検索することが可能となる。

また、距離テーブルを用いることにより、複雑な距離計算などを行う必要もなく高速な検索が可能となる。

また、距離テーブルを用いることにより、誤認識の許容度合いを適切な値に設定することが可能となり、効率のよい検索が可能となる。

さらに、検索パラメータ (例えば、文字認識の信頼度) を文書データに付与しておくことにより、文書データや文字列要素に合わせた検索のための基準の選択や検索手法の切り替えが可能となり、検索の精度を向上させることが可能となる。

#### (実施の形態 2)

実施の形態 2 では、複数の距離テーブルが用意される。これは、文字要素と文字要素との間に、文字要素と文字要素との類似度に関連する複数の距離が予め設定されていることを意味する。その複数の距離テーブルのうちの 1 つが選択され、その選択された距離テーブルが文字要素列を検索するために使用される。これは、文字要素間に予め設定されている複数の距離のうちの 1 つを選択し、その選択された距離と所定の基準距離との比較結果に基づいて、文字要素の一致、不一致を判定することを意味する。文字要素間に予め設定されている複数の距離のうちの 1 つは、例えば、ユーザからの入力に基づいて選択され得る。

なお、文字要素列を検索する検索処理の基本的な流れは実施の形態 1 と同様である。

複数の距離テーブルは、例えば、複数種類の文字認識システムにそれぞれ対応するように用意される。あるいは、複数の距離テーブルは、複数の文字種 (例え

ば、漢字、アルファベット、ギリシャ文字、カタカナなど）にそれぞれ対応するように用意されてもよいし、複数のフォント種にそれぞれ対応するように用意されてもよい。

5 例えば、図2は、「ゴシック体」というフォント種に対応するように用意された距離テーブルの例を示し、図7は、「明朝体」というフォント種に対応するように用意された距離テーブルの例を示す。

10 検索対象となる文書データに応じて、複数の距離テーブルのうちの1つが選択的に使用される。例えば、オリジナル文書を文字認識することによって得られる文書データに、その文書データに含まれる文字種、フォント種、文字認識に使用された文字認識システムの種類などの情報を検索パラメータとして保持しておくことにより、検索の際に適切なテーブルを選択して用いることが可能となる。その結果、検索の精度と速度を向上させることが可能となる。

15 フォント種に応じて距離テーブルを切り替える場合には、オリジナル文書を文字認識することによって得られる文書データに、フォント種が明朝体に近いかゴシック体に近いかを示す情報を文字要素ごとに検索パラメータとして付与しておくことが好ましい。この検索パラメータを参照することにより、ゴシック体の文字要素を含む文書データから文字要素列を検索する場合には図2に示されるような距離テーブルを用い、明朝体の文字要素を含む文書データから文字要素列を検索する場合には図7に示されるような距離テーブルを用いることが可能になる。

20 フォント種を示す情報は、文字認識を行うと同時にフォントの種類を認識することなどにより得ることができる。

25 また、同一の文書データに対して複数の距離テーブルを切り替えるようにしてもよい。この場合、特定の距離テーブルを用いて検索できなかった文書データについて再度、検索したい文字要素列の有無を異なる尺度で検証することができる。その結果、検索の精度を向上させることが可能となる。更に、距離テーブルを用いた検索の結果に対して、その検索で得られた文字要素列の位置に対応する文書

の画像イメージを用いて再度高精度な文字認識を行ってもよい。これにより、距離テーブルを用いて高速な粗検索を行い候補を絞った後に、高精度な文字認識（一般的に処理時間がかかる）を用いて検索対象を確定することが可能となり、検索精度と検索速度の両立が可能となる。

5       特に、文字数の少ない検索文字列（2文字単語など）を検索する場合、類似した文字列が偶然検索される可能性が高い。よってこのような場合にも指定した検索文字列の文字要素数に基づいて異なるテーブルを用いて再検証したり、高精度な文字認識を併用したりすることで、処理時間を必要以上に増さずに、精度の高い検索が可能となる。

10       また、検索キーワードとして指定された文字要素列の字種に応じて、距離テーブルを切り替えるようにしてもよい。例えば、検索キーワードとして指定された文字要素列および文書データがアルファベットのみを含む場合には、アルファベット用の距離テーブルを用いることにより、余分な検索処理を省くことが可能となる。

15       なお、上述した実施の形態では、1文字単位の文字認識の誤りを補うために距離テーブルを用いていた。しかし、1文字単位以上の文字認識の誤りを補うために距離テーブルを用いることも可能である。

20       図8Aは、複数の文字要素を単一の文字要素として扱った結果生じる誤認識の例を示す。図8Aには、2つの「木（き）」を「林（はやし）」と誤認識した例と、2つの「0（ゼロ）」を「∞（無限大）」と誤認識した例とが示されている。

      図8Bは、単一の文字要素を複数の文字要素として扱った結果生じる誤認識の例を示す。図8Bには、「川」を3つの「1（いち）」と誤認識した例と、「い」を「し」と「1（いち）」と誤認識した例とが示されている。

25       距離テーブルにおいて、「木（き）」2文字と「林」との距離、「0（ゼロ）」2文字と「∞（無限大）」との距離、「川」と3つの「1（いち）」との距離、「い」と「し」、「1（いち）」との距離をそれぞれ小さい値に設定して



おくことにより、図 8 A および図 8 B に示されるように誤認識されている場合でも正しい検索結果を得ることができる。

図 9 は、図 8 A および図 8 B に示される文字要素を含む文字要素間の距離テーブルの一例を示す。

5 図 9 に示される例では、「木 (き)」2 文字と「林」との距離、「0 (ゼロ)」2 文字と「 $\infty$  (無限大)」との距離、「川」と3つの「1 (いち)」との距離、「い」と「し」、「1 (いち)」との距離は、それぞれ、13 以下に設定されている。これらの距離は、文字要素の他の組み合わせに対応する距離 (98 以上) よりもかなり小さい値に設定されている。

10 ここで、検索キーワードとして文字要素列「100」を指定し、誤りの度合いの基準となる距離 (基準距離) を50とすると、「1」を検索した後、「0」を2つ検索するとともに「 $\infty$ 」を検索することが可能になる。これにより、「100」が「1 $\infty$ 」と誤認識された場合でも、検索キーワードとしての文字要素列「100」を検索することが可能となる。

15 同様に、「いろり」が「し1ろり」と誤認識された場合でも、検索キーワードとして文字要素列「いろり」を検索することが可能になる。

更に、かな漢字変換等の誤りによってオリジナルの文書自体に文章として誤った表現が含まれる場合 (「納める」を「収める」と表現) や、複数の送り仮名付け方が存在する場合 (「変る」を「変わる」) や、漢字表記した言葉をひらがなで検索しようとする場合 (「切磋」を「せっさ」で検索) や、類義語で検索しようとした場合 (「価格」を「定価」で検索) や、異なる言語に対して検索しようとした場合 (「history」を「歴史」で検索する場合) についても、文字要素間の距離テーブルにおいてそれぞれ「収」と「納」との距離、「変わ」と「変」との距離、「切磋」と「せっさ」との距離、「価格」と「定価」との距離、  
20 「history」と「歴史」との距離をそれぞれ小さい値に設定しておくことにより、正しい検索結果を得ることができる。

(実施の形態 3)

実施の形態 3 では、文字要素間の距離テーブルにおいて、文字要素間の距離に加えて文字要素の発生頻度が予め設定される。これにより、文字要素間の距離が確率的な分布を有するように距離を扱うことが可能になる。

5 図 10 A は、発生頻度付きの距離テーブルの例を示す。図 10 A は、文字要素「下」に対して、文字要素「T」の発生頻度（確率）が距離 10 では 0.2、距離 20 では 0.6、距離 30 では 0.2であることを示す。

10 図 10 B は、発生頻度付きの距離テーブルの他の例を示す。図 10 B に示される例では、発生頻度は正規分布するという仮定の下に、発生頻度が、距離と分散とによって表されている。例えば、図 10 B は、文字要素「下」に対して、文字要素「T」は距離 20 を中心とし分散 10 の正規分布に位置していることを表している。

15 図 10 C は、発生頻度付きの距離テーブルの他の例を示す。図 10 C に示される例では、発生頻度は一様分布するという仮定の下に、発生頻度が、その分布の最短距離と最大距離とによって表されている。例えば、文字要素「下」に対して、文字要素「F」は距離 50 から距離 70 まで一様分布し、文字要素「ト」は距離 63 から距離 122 まで一様分布している。従って、距離 63 から距離 70 までの間には文字要素「F」と文字要素「ト」とが重複して分布するのでそれぞれの文字要素の発生頻度は 0.5 であると判断される。

20 このように、図 10 A ~ 図 10 C に示されるように距離テーブルに発生頻度を付与することにより、文字要素と文字要素との距離を固定した 1 つの値でなく、確率的な分布を有する幅を持った値として取り扱うことができる。これにより、文字要素の発生頻度に応じた検索を行うことが可能となる。

25 以下、図 11 を参照して、発生頻度付きの距離テーブルを用いて文字要素列を検索する検索処理を説明する。この検索処理の基本的な流れは、実施の形態 1、2 と同様である。

まず、文字認識の信頼度に応じて検索時に許容される文字要素間の距離（基準距離）が決定される。その後、その基準距離に対応する文字要素の発生頻度が距離テーブルから算出される。ここで、文字認識の信頼度、基準距離および発生頻度の関係は予め設定されているものとする。

図 1 1 は、文字認識の信頼度、基準距離および発生頻度の関係の一例を示す。文書データ中の文字要素「人」（信頼度 0.9）に対して許容される距離（基準距離）を 10 とすると、検索キーワードとしての文字要素「人」に対して文書データ中の文字要素「人」が対応する割合（発生頻度）は 0.9 となっている。同様に、文書データ中の文字要素「区」（信頼度 0.4）に対して許容される距離（基準距離）を 60 とすると、検索キーワードとしての文字要素「口」に対して文書データ中の文字要素「区」が対応する割合（発生頻度）は 0.1 となる。文字要素「構」・「成」についても同様に発生頻度は 0.9 となる。

ここで、検索キーワードとして指定された文字要素列「人口構成」を構成する各文字要素について文書データ中の文字要素列と対応する割合が平均 0.7（ $= (0.9 + 0.1 + 0.9 + 0.9) / 4$ ）となる。あらかじめ検索の基準として一致する割合を例えば平均値で 0.5 以上としておくことで、上記の誤認識された文字要素列「人区構成」を検索キーワードとしての文字要素列「人口構成」から検索することが可能となる。

更に、検索の結果を表示する際に、一致する割合に応じて表示を切り替えることも可能である。例えば、一致する割合の高さに応じて文書画像中の対応する位置に強調表示（輝度や色、点滅などによる強調）を行うことにより、検索者が一致割合を視覚的に確認することが容易となる。

また、上述した例では、検索キーワードを構成する文字要素の一致する割合の平均値を検索の基準としたが、最低値を基準としてもよいし、高い一致割合（例えば 0.8 以上）の文字要素が文字要素列全体のある割合以上（例えば半分以上）を占めている場合を基準としてもよい。更に、高い一致割合（例えば 0.8

以上)の文字要素が文字要素列全体のある割合以上(例えば $2/3$ 以上)あれば、残りの一致割合の低い文字要素に対しては検索の基準をゆるくして検出しやすいように変更してもよい。

例えば、図12に示されるように、「人口構成」が誤認識された「人同構成」という文字要素列を含む文書データから、検索キーワードとしての文字要素列「人口構成」を検索することを考える。ここで、「口」が誤認識された「同」は文字認識の信頼度が0.3で、許容される距離(基準距離)が80となり、「口」に一致する割合は0.0となっている。このままでは、「同」と「口」とは全く一致しない扱いとなる。しかし、検索キーワードを構成する他の文字要素「人」・「構」・「成」については一致割合が高い(0.9)ので「同」に対しては許容される距離(基準距離)を80よりも大きい値(例えば120)に設定することにより、検索キーワード「人口構成」に対応する文字要素列を検出することができるようになる(図6に示される距離テーブルを使用する場合)。

図13は、発生頻度付きの距離テーブルを作成する処理の手順を示す。図13は、文字要素「X」と文字要素「Y」との間の距離と発生頻度とをどのように定義するかを示す。文字要素間のすべての組み合わせに対して同様の処理を行うことにより、文字要素間のすべての組み合わせに対して距離と発生頻度とを定義することができる。

図13に示される反復処理を統計的に十分な回数だけ行うことにより、文字要素「X」と文字要素「Y」との距離がDとなる発生頻度を統計的に確からしい値として得ることが可能になる。

なお、発生頻度付きの距離テーブルを作成するために使用されるニューラルネットワークNNは、文字を予め学習させているものとする。ニューラルネットワークNNのタイプは問わない。

図14は、発生頻度付きの距離テーブルを用いて、認識結果の中から検索キーワードに対応する文字要素列を検索する検索処理の手順を示す。この検索処理は、

CPU 110によって文書検索プログラムに従って実行される。

図14に示される検索処理によって、検索キーワードの文字要素列に含まれる文字要素「X」に対応する文字要素が認識結果に含まれているか否かが判定される。検索キーワードの文字要素列に含まれるすべての文字要素について同様の検索処理が繰り返される。

検索キーワードの文字要素列に含まれるすべての文字要素を0より大きい発生頻度で連続して認識結果から検出することができた場合には、その検索キーワードの文字要素列に対応する発生頻度の列が得られる。この発生頻度の列に基づいて、検索キーワードの文字要素列に対応する文字要素列が認識結果に含まれているか否かが判定される。この判定は、発生頻度の列の平均値に基づいて行われてもよいし、発生頻度の列の最低値に基づいて行われてもよい。

図14において、信頼度Rは文字認識の信頼度を示す。信頼度Rは、認識結果に含まれる各文字要素に対して予め設定されている。また、信頼度Rと基準距離Dとの関係も予め設定されている。

このように、距離テーブルに、文字要素間の距離に加えて、発生頻度の情報を付与しておくことにより、同じ距離でも発生頻度に応じて検索の基準や手続きを切り替えることが可能となる。その結果、より精度の高い検索が可能となる。

(実施の形態4)

実施の形態4では、文字要素間の距離テーブルを用いて、検索キーワードとして指定された文字要素列が予め複数の文字要素列に展開される。複数の文字要素列のそれぞれについて検索処理が実行される。

以下、「日本」という文字要素列を検索キーワードとして指定して、図4に示される認識結果の中から検索キーワードに対応する文字要素列を検索する検索処理を説明する。この検索処理は、CPU 110によって文書検索プログラムに従って実行される。

最初に、文字要素列「日本」が文字要素「日」と文字要素「本」とに分割され

る。各文字要素について文字要素間の距離テーブルが参照され、各文字要素と、各文字要素との距離が所定の基準距離（例えば、150）よりも小さい文字要素（例えば「日」については「日」と「目」、「本」については例えば「本」と「木」と「大」）とが組み合わせられる。その結果、検索キーワードとして指定された文字要素列「日本」に基づいて、新たな文字要素列「日本」、「目本」、「日木」、「目木」、「日大」、「目大」が生成される。

次に、図4に示される認識結果（文書データ）から新たな文字要素列のそれぞれを検索する検索処理が実行される。これにより、新たな文字要素列「日木」をオリジナル文書の「日本」が存在する位置において検出することが可能になる。その結果、望ましい検索結果を得ることができる。

ここで、検索した結果何も検出されない場合には、所定の基準距離をより大きい値（例えば、200）に再設定し、新たにより多くの文字要素列を生成して同様の検索処理を実行するようにしてもよい。これにより、距離テーブルにおいて許容される距離（基準距離）が150の場合には検出することができなかった認識誤りを検出することができるようになる。

このように、文字要素間の距離テーブルを用いて、検索キーワードとして指定された文字要素列を誤りの可能性のある複数の文字要素列に置き換え、その複数の文字要素列のそれぞれについて検索処理を実行することにより、実施の形態1と同様に、指定された文字要素列が誤認識で他の文字要素列に置き換わったような文字要素列を文書データから検索することが可能となる。

距離テーブルを用いることにより、複雑な距離計算などを逐次行う必要がなく高速な検索が可能となる。

また、距離テーブルを用いることにより、誤認識の許容度合いを適切な値に設定することが可能となり、効率のよい検索が可能となる。

（実施の形態5）

実施の形態5では、文字要素間の距離テーブルを用いて、認識結果（文書デー

タ)に含まれる各文字要素に1以上の他の文字要素が追加された後に、検索処理が実行される。

以下、「日本」という文字要素列を検索キーワードとして指定して、図4に示される認識結果の中から検索キーワードに対応する文字要素列を検索する検索処理を説明する。この検索処理は、CPU110によって文書検索プログラムに従って実行される。

はじめに、文字要素間の距離テーブルが参照され、認識結果(文書データ)に含まれる各文字要素との距離が所定の基準距離(例えば、150)よりも小さい文字要素(例えば「日」については「日」と「目」、「木」については「本」と「大」など)が認識結果の候補として認識結果に追加される。

図15は、認識結果の候補が追加された例を示す。図15に示される例では、認識結果の文字要素「日」に対して認識結果の候補として文字要素「目」が追加されている。また、認識結果の文字要素「木」に対して認識結果の候補として文字要素「本」と文字要素「大」とが追加されている。

次に、検索キーワードとして指定された文字要素列「日本」が文字要素「日」と文字要素「本」とに分割される。文字要素「日」について検索処理を実行することにより、図15に示される認識結果(文書データ)から文字要素「日」が検出される。次に、文字要素「日」が検出された位置の次の文字要素(「木」、「本」、「大」)の中に文字要素「本」が存在するか否かが判定される。文字要素(「木」、「本」、「大」)の中に文字要素「本」が含まれているので、認識結果から文字要素列「日本」が検出されたと判定される。

検索キーワードとして指定された文字要素列に含まれる文字要素の数が3つ以上である場合にも、同様の手順に従って、認識結果から文字要素列を検出することができる。すなわち、検索キーワードとして指定された文字要素列に含まれるすべての文字要素が連続して認識結果から検出された場合に指定された文字要素列が認識結果から検出されたと判定すればよい。

このように、文字要素間の距離テーブルを用いて、認識結果（文書データ）に含まれる各文字要素に1以上の文字要素を予め追加しておくことにより、実施の形態1と同様に、指定された文字要素列が誤認識で他の文字要素列に置き換わった場合でも、指定された文字列要素に対応する文字要素列を文書データから検索

することが可能となる。

また、認識結果（文書データ）に含まれる各文字要素に1以上の文字要素を予め追加しておくことにより、検索時に距離テーブルを参照する手続きを省略することが可能となる。

なお、実施の形態5において、実施の形態1～3で説明した検索の過程で距離テーブルを用いる形態や、実施の形態4で説明した距離テーブルを用いて検索キーワードとして指定された文字要素列を複数の文字要素列に展開する形態を併用することも可能である。

#### （実施の形態6）

図16は、段落A～段落Dを有するオリジナル文書の例を示す。図16に示される例では、段落の正しい順序は、段落A、段落B、段落C、段落Dの順番であると仮定する。すなわち、段落Aの末尾の文章は段落Bの先頭の文章に続き、段落Bの末尾の文章は段落Cの先頭の文章に続き、段落Cの末尾の文章は段落Dの先頭の文章に続くとは仮定する。

図16に示されるオリジナル文書を文字認識することによって文書データ（認識結果）を得る場合、その文書データ（認識結果）においてオリジナル文書の段落間の接続関係が正しく認識されるとは限らない。段落のレイアウトの仕方は文書によって様々であるため、段落間の接続関係を自動的に認識することは極めて困難だからである。従って、段落間の接続関係の認識誤りが発生し得る。

実施の形態6は、段落間の接続関係の認識誤りが発生した場合でも、認識結果から文字要素列を正しく検索することが可能な検索方法を提供する。

図17は、段落間の接続関係の認識結果を示すテーブルの例を示す。このよう



なテーブルは、CPU 110が文書登録プログラムを実行することによって生成され、メモリ 140に格納される。

認識された各段落は、各段落に固有の段落番号（図 17では、A、B、C、D）によって識別される。

5 図 17に示されるテーブルは、各段落について、次に接続される可能性のある段落の番号と、段落単位の認識結果とを有している。例えば、図 17に示されるテーブルの第 1行は、段落Aに接続される可能性のある段落が段落Bまたは段落Cのいずれかであり、段落Aの末尾が「日本の」であることを示している。

10 ここで、特定の段落に接続される可能性のある段落は、例えば、オリジナル文書の画像データを文字認識する際に各段落同士のそれぞれの位置関係を参照することによって決定される。例えば、オリジナル文書の文章が縦書きの場合には、ある段落Xの次に続く可能性のある段落は、段落Xよりも下に位置する段落（例えば、段落Y）、または、段落Xよりも左に位置する段落（例えば、段落Z）のいずれかであると決定される。この場合には、段落Xに接続される可能性のある段落として、段落Yと段落Zとが上述したテーブルに登録される。

15 さらに、文字認識した結果に基づいて、ある段落Xの末尾の文章と文法的に接続可能な文頭を有する段落を、段落Xに接続される可能性のある段落として決定してもよい。

20 また、オリジナル文書の段落が特定の規則に従ってレイアウトされている場合には、その特定の規則に基づいて、段落Xに接続される可能性のある段落を決定してもよい。

以下、図 17に示されるテーブルを用いて、認識結果から文字要素列「日本の人口」を検索する検索処理を説明する。この検索処理は、CPU 110によって文書検索プログラムに従って実行される。

25 実施の形態 1～実施の形態 5と同様に、文字要素列「日本の人口」を構成する各文字要素「日」、「本」、「の」、「人」、「口」が各段落から順次検索され

る。

図 1 7 に示される例では、段落 A の末尾に文字要素「日」、「本」、「の」が続いて検出される。次に、文字要素「人」が検索される。ここで、段落 A に接続される可能性のある段落は、段落 B または段落 C であるため、段落 B と段落 C のそれぞれの先頭に文字要素「人」が存在するか否かが判定される。図 1 7 に示される例では、段落 B の先頭において文字要素「人」が検出されるため、その次の位置に文字要素「口」が存在するか否かが判定される。その結果、最終的に、文字要素列「日本の人口」を構成するすべての文字要素が検出されることになる。

図 1 8 は、図 1 7 に示されるテーブルと文字要素間の距離テーブル（例えば、図 2）とを用いて、段落間の接続関係を考慮して、認識結果から検索キーワードの文字要素列を検索する検索処理の手順を示す。この検索処理は、CPU 110 によって文書検索プログラムに従って実行される。

図 1 8 に示される検索処理によって、検索キーワードの文字要素列に含まれる文字要素「X」に対応する文字要素が認識結果に含まれているか否かが判定される。検索キーワードの文字要素列に含まれるすべての文字要素について同様の検索処理が繰り返される。

文字要素「X」と段落 A の末尾の文字要素とが一致しなかったと仮定する。この場合には、文字要素「X」と段落 A に接続される可能性のある段落 B または段落 C のうちの 1 つの段落の先頭の文字要素「Y」とが一致するか否かが判定される。段落 A に接続される可能性のある段落が段落 B または段落 C であることは、図 1 7 に示されるテーブルにおいて予め定義されている。

文字要素「X」と文字要素「Y」とが一致するか否かは、文字要素間の距離テーブルを用いて判定される。この判定の方法は、実施の形態 1 ～実施の形態 5 で説明したとおりである。

このように、特定の段落に接続される可能性のある複数の段落を予め定義しておくことにより、文字認識の際に段落の接続関係が誤認識された場合でも、複数

の段落にまたがる文字要素列を正しく検出することが可能となる。

また、段落同士の接続に限らず、段落内で行と行との接続があいまいな場合（例えば、行間に図、表、見出しなどが挿入されている場合）にも同様に行ごとに異なる番号を付与し、特定の行に接続される可能性のある複数の行を予め定義しておくことにより、複数の行にまたがる文字要素列を正しく検出することが可能となる。

また、文字要素と文字要素との接続があいまいな場合（例えば、文字要素間に図、表などが挿入されている場合や、文字要素列の配置が装飾的な場合〔曲線状に配置された文字要素列〕など）にも同様に行（文字要素）ごとに異なる番号を付与し、特定の行（文字要素）に接続される可能性のある複数の行（文字要素）を予め定義しておくことにより、複数の行（文字要素）にまたがる文字要素列を正しく検出することが可能となる。

このように、認識結果に含まれる複数の文字要素のうち特定の文字要素に対して、その特定の文字要素に接続される可能性のある複数の文字要素が予め決定される。検索処理では、その特定の文字要素と、その特定の文字要素に接続される可能性のある複数の文字要素のうちの1つとを接続することによって得られる文字要素列が、検索キーワードの文字要素列の少なくとも一部に一致するか否かが判定される。これにより、文字認識の際に文字要素の接続関係が誤認識された場合でも、複数の文字要素にまたがる文字要素列を正しく検出することが可能となる。

その特定の文字要素は、行または列の末尾に配置されており、その特定の文字要素に接続される可能性のある複数の文字要素のそれぞれは、行または列の先頭に配置されていてもよい。

さらに、その特定の文字要素と、その特定の文字要素に接続される可能性のある複数の文字要素のうちの1つとは、同一の行または列に配置されており、その特定の文字要素と、その特定の文字要素に接続される可能性のある複数の文字要素

素のうちの他の1つとは、異なる行または列に配置されており、かつ、同一の列または行に配置されていてもよい。

なお、上述した例では、特定の段落（行または文字要素）の後に接続される可能性のある複数の段落（行または文字要素）を予め定義しておく例を説明した。

5     あるいは、特定の段落（行または文字要素）の前に接続される可能性のある複数の段落（行または文字要素）を予め定義しておくようにしてもよい。この場合でも上述した効果と同様の効果が得られる。

10     また、特定の段落（行または文字要素）に接続される可能性のある段落（行または文字要素）の番号は、上述したように、段落（行または文字要素）番号の絶対値を用いて表現されてもよいし、段落（行または文字要素）番号の相対値を用いて表現されてもよい。例えば、段落Aに接続される可能性のある段落を段落B、段落Cと表現する代わりに、段落+1、段落+2と表現するようにしてもよい。

（実施の形態7）

15     図19は、段落間の接続関係の認識結果を示すテーブルの例を示す。このようなテーブルは、CPU110が文書登録プログラムを実行することによって生成され、メモリ140に格納される。

20     図19に示されるテーブルは、各段落について、段落単位の認識結果と、段落単位の位置を有している。段落単位の位置は、例えば、オリジナル文書の右上隅を原点するXY座標で表される。X軸方向、Y軸方向は、例えば、図16に示される方向である。例えば、図19に示されるテーブルの第1行は、段落Aの末尾が「日本の」であり、段落Aの位置が座標(X, Y) = (10, 100)によって表されることを示している。

25     検索処理の手順は、実施の形態6とほぼ同様である。段落Aの末尾において文字要素「日」、「本」、「の」が続いて検出された後に、文字要素「人」を検索する際、段落Aに接続される可能性のある段落が、図19に示されるテーブルに格納されている段落の位置を示す座標に基づいて決定される。この場合、段落A

の座標は  $(X, Y) = (10, 100)$  である。段落Aに接続される可能性のある段落として、段落AのX座標と等しいX座標を有し、かつ、段落AのY座標の次に大きいY座標を有する段落C  $(X, Y) = (10, 200)$  と、段落AのY座標と等しいY座標を有し、かつ、段落AのX座標の次に大きいX座標を有する段落B  $(X, Y) = (100, 100)$  とが決定される。

この場合、段落Bの先頭において文字要素「人」が検出されるので、その次の位置に文字要素「口」が存在するか否かが判定される。その結果、最終的に、文字要素列「日本の人口」を構成するすべての文字要素が検出されることになる。

なお、オリジナル文書の文章が縦書きの場合には、ある段落Xの次に続く可能性のある段落を、段落Xよりも下に位置する段落（例えば、段落Y）、または、段落Xよりも左に位置する段落（例えば、段落Z）のいずれかであると決定してもよい。

また、オリジナル文書の段落が特定の規則に従ってレイアウトされている場合には、その特定の規則に基づいて、段落Xに接続される可能性のある段落を決定してもよい。

なお、XY座標の原点やX座標軸、Y座標軸の方向は自由に選んでよく、また座標値についても段落や図ごとに番号を割り振った値の順番を座標値の単位として用いてもよい。

このように、段落ごとに段落の位置を示す情報を保持しておくことにより、文字認識の際に段落の接続関係を誤認識した場合にでも、複数の段落にまたがる文字要素列を検出することが可能となる。

また、段落の位置を示す座標を保持しておくことにより、文書データを変更することなく特定の段落に接続される可能性のある段落の決定方法を変更することが可能であり、段落の位置を示す座標は文書のレイアウトを再現するために用いることも可能である。

なお、上術した例では段落ごとに段落の位置を示す座標を保持することとした

が、行単位や文字要素単位に異なる番号を付与し、行ごとにまたは文字要素ごとにその位置を示す座標を保持するようにしてもよい。これにより、複数の行または複数の文字要素にまたがる文字要素列を検索することが可能となる。

(実施の形態 8)

- 5        実施の形態 6 と同様に、図 16 に示されるオリジナル文書を文字認識することによって得られる文書データ（認識結果）から文字要素列「日本の人口」を検索する検索処理を説明する。

10        ここで、認識結果は、図 20 に示されるように、特定の段落と、その特定の段落に接続される可能性のある複数の段落のうちの 1 つとを接続した形式で保持される。このような認識結果は、例えば、メモリ 140 に格納される。

15        図 20 に示される例では、2 種類の文字認識結果（文字認識結果 1、文字認識結果 2）が保持されている。文字認識結果 1 は、段落 A と、段落 A に接続される可能性のある段落 C とを接続することによって得られる。文字認識結果 2 は、段落 A と、段落 A に接続される可能性のある段落 B とを接続することによって得られる。

      図 20 に示される認識結果から文字要素列「日本の人口」を検索する場合には、文字認識結果 1 と文字認識結果 2 のそれぞれに対して「日本の人口」が検索される。その結果、文字認識結果 2 から「日本の人口」を検出することができる。

20        なお、図 20 に示されるように複数の段落の接続関係を想定して認識結果を保持する場合には、検索キーワードとして指定される文字要素列に含まれる文字要素の数に上限値（例えば、10 文字要素）を設け、段落 A に接続される可能性のある段落 B、段落 C の先頭から 9 文字要素のみを段落 A の認識結果に付加して保持するようにしてもよい。この場合、段落 A から段落 B または段落 C にまたがる 10 文字要素までの文字要素列の検索が可能となる。

25        このように、特定の段落に接続される可能性のある複数の段落を考慮して、複数の認識結果を予め保持しておくことにより、文字認識の際に段落の接続関係を

誤認識した場合でも、複数の段落にまたがる文字要素列を検出することが可能となる。また、複数の認識結果を予め保持しておくことにより、検索処理の手順が簡易になり、従来の検索処理を利用することが可能となる。

(実施の形態 9)

5        図 2 1 に示されるようなレイアウトを有するオリジナル文書を文字認識することによって得られる文書データ（認識結果）から文字要素列「神戸」を検索する検索処理を説明する。

10        図 2 1 に示される例では、オリジナル文書の文章は横書きであるが、文字要素同士の間隔が縦方向に接近しているため、文字認識を行った場合に文字要素間の接続関係が誤認識される可能性がある。

      実施の形態 9 は、文字要素間の接続関係が誤認識された場合でも、認識結果から文字要素列を正しく検索することが可能な検索方法を提供する。

15        図 2 2 A は、オリジナル文書の文章が縦書きであるという仮定の下で文字認識を実行した場合の認識結果を示すテーブルの例を示す。これは、図 2 1 に示されるオリジナル文書を列単位に認識した結果である。

      図 2 2 B は、オリジナル文書の文章が横書きであるという仮定の下で文字認識を実行した場合の認識結果を示すテーブルの例を示す。これは、図 2 1 に示されるオリジナル文書を行単位に認識した結果である。

20        このようなテーブルは、CPU 1 1 0 が文書登録プログラムを実行することによって生成され、メモリ 1 4 0 に格納される。

25        検索キーワードとして文字要素列「神戸」が指定された場合には、図 2 2 A に示される認識結果から文字要素列「神戸」が検索され、かつ、図 2 2 B に示される認識結果から文字要素列「神戸」が検索される。その結果、図 2 2 B に示されるテーブルの行番号 3 に対応する認識結果から文字要素列「神戸」が検出される。これにより、図 2 1 に示されるオリジナル文書に文字列「神戸」が含まれていたことが分かる。

このように、複数のレイアウトに対応した認識結果を保持しておくことにより、正しいレイアウトを認識することが困難な文書に対しても、認識結果から文字要素列を検索することが可能となる。また、複数のレイアウトに対応した認識結果を保持しておくことにより、従来の検索処理を利用することが可能となる。

5       なお、上述した例では、縦書き、横書きの2種類のレイアウトを想定したが、レイアウトの種類はこれらに限定されない。例えば、縦書き、横書き以外に、斜め方向のレイアウトなどその他のレイアウトを同様に扱うことが可能である。

（実施の形態10）

10       実施の形態10は、オリジナル文書のレイアウトが誤って認識された場合でも、認識結果から指定された文字要素列を正しく検索することが可能な検索方法を提供する。

図21のようなレイアウトを有するオリジナル文書を文字認識することによって得られる文書データ（認識結果）から文字要素列「神戸」を検索する検索処理を説明する。

15       ここで、オリジナル文書のレイアウトを誤って認識した例として、図22Aに示されるテーブルが保持されていると仮定する。実施の形態9とは異なり、図22Bに示されるテーブルは保持されていない。

20       はじめに、文字要素列「神戸」が文字要素「神」と文字要素「戸」とに分割される。図22Aに示されるテーブルを用いて、各文字要素について検索処理が行われる。その結果、図22Aに示されるテーブルの列番号5に対応する認識結果の第3文字目から文字要素「神」が検出され、図22Aに示されるテーブルの列番号4に対応する認識結果の第3文字目から文字要素「戸」が検出される。

25       文字要素列「神戸」を構成するすべての文字要素が認識結果から検出された場合には、各文字要素が検出された位置関係に基づいて、認識結果から文字要素列「神戸」が検出されたか否かが判定される。この場合、文字要素「神」と文字要素「戸」とは、隣接する行の同じ文字数目に連続して検出されたため、文字要素



列「神戸」が検出されたと判定される。

なお、文字要素列が検出されたか否かの判定には、上述した基準と異なる基準を使用してもよい。すなわち、各文字要素が検出された位置関係が上述した関係とは異なる関係を満たす場合に、文字要素列が検出されたと判定するようにしてもよい。例えば、文字要素の位置を示す座標が分かっている場合には、個々の文字要素があらかじめ定めた距離以下で接近しかつ直線的に配置されている場合に、文字要素列が検出されたと判定するようにしてもよい。

また、検索処理の手順として他の手順を使用することもできる。例えば、上述したように文字要素をすべて検索するのではなく、文字要素「神」を検出できた場合にのみ文字要素「神」が検出されや行に隣接する行からのみ文字要素「戸」を検索するようにしてもよい。これにより、検索処理のうち不要な部分を削減することができるので、検索処理を効率的に行うことが可能となる。

このように、検索キーワードとして指定された文字要素列に含まれる各文字要素を認識結果から検出し、各文字要素の検出された位置関係に基づいて文字要素列が検出されたか否かが判定される。これにより、オリジナル文書のレイアウトが誤って認識された場合でも、認識結果から指定された文字要素列を正しく検索することが可能となる。

#### (実施の形態 11)

実施の形態 11 では、検索キーワードとして指定された文字要素列が 2 以上の文字要素列に分割され、その分割された文字要素列が検出された段落間の位置関係に基づいて、検索キーワードが検出されたか否かが判定される。

以下、図 16 に示されるレイアウトを有するオリジナル文書を文字認識することによって得られる文書データ（認識結果）から文字要素列「日本の人口」を検索する検索処理を説明する。

実施の形態 7 と同様に、段落間の接続関係の認識結果を示すテーブル（図 19）が予め用意される。

図19に示されるテーブルは、各段落について、段落単位の認識結果と、段落単位の位置を有している。段落単位の位置は、例えば、オリジナル文書の右上隅を原点するXY座標で表される。X軸方向、Y軸方向は、例えば、図16に示される方向である。

5       はじめに、検索キーワードとして指定された文字要素列「日本の人口」が2つの文字要素列に分割される。例えば、文字要素列「日本の人口」は、文字要素列「日本」と文字要素列「の人口」に分割される。

10       次に、分割された2つの文字要素列のそれぞれを個々の段落から検索する。文字要素列「日本の人口」のすべての分割の仕方について検索が繰り返される。例えば、文字要素列「日本の人口」が文字要素列「日本の」と文字要素列「人口」とに分割された場合には、図16に示される例では、段落Aの末尾から文字要素列「日本の」が検出され、段落Bの先頭から文字要素列「人口」が検出される。分割された文字要素列がすべて検出された場合には、検出された段落間の位置関係に基づいて、文字要素列「日本の人口」が検出されたか否かが判定される。

15       例えば、2つの文字要素列が検出された段落が隣接していたり、2つの文字要素列が検出された段落の位置が近い場合には、検索キーワードとして指定された文字要素列が検出されたと判定される。図16に示される例では、文字要素列「日本の」が検出された段落Aの座標 $(X, Y) = (10, 100)$ と、文字要素列「人口」が検出された段落Bの座標 $(X, Y) = (100, 100)$ とは、  
20       同一のY座標を有し、かつ、互いに隣接する。従って、認識結果から文字要素列「日本の人口」が検出されたと判定される。

      また、上述したように文字要素列を分割することによって得られる文字要素列を段落から検索する場合には、各段落の文末と先頭のみに対して検索処理を行うことが好ましい。これにより、検索処理の効率を向上させることができる。

25       なお、上述した例では、検索キーワードとして指定された文字要素列を2つの文字要素列に分割した例を説明したが、検索キーワードとして指定された文字要

素列を必要に応じて3つ以上の文字要素列に分割することも可能である。この場合にも、同様の検索処理を行うことが可能である。

また、上述した例では、複数の段落にまたがる文字要素列を検索する例を示したが、同様にして、複数の行にまたがる文字要素列を検索することも可能である。  
5 この場合には、検索キーワードとして指定された文字要素列を分割し、各行に対して分割された文字要素列のそれぞれを検索し、分割されたすべての文字要素列が隣接した位置に検出された場合に検索キーワードが検出されたと判定すればよい。

10 このように、実施の形態6～実施の形態11では、段落や行の接続が誤っている（または不定な）場合にでも、複数の段落にまたがる文字要素列を正しく検出することが可能となる。

また、文字認識の誤りがある場合や、段落間・行間の接続が誤っていたり不定な場合や、縦書き・横書きの判断が誤っているあるいは不定である場合に指定され文字要素列を検索することが可能である。

15 なお、実施の形態1～実施の形態11のそれぞれを単独で実施することも可能であるし、それらの少なくとも2つを組み合わせることも可能である。

本発明の検索処理は、典型的には、コンピュータ上のソフトウェアによって実現される。しかし、本発明の検索処理をハードウェアによって実現してもよいし、ソフトウェアとハードウェアとの組み合わせによって実現してもよい。

20 本発明の検索処理の一部または全部を表現するプログラム（文書検索プログラム）は、例えば、メモリ170に格納されている。あるいは、文書検索プログラムは、フロッピーディスク、CD-ROM、DVD-ROMなどの任意のタイプの記録媒体に記録され得る。そのような記録媒体に記録された文書検索プログラムは、ディスクドライブ（図示せず）を介してコンピュータのメモリにロードされる。  
25 あるいは、文書検索プログラム（またはその一部）は、通信網（ネットワーク）または放送を通じてコンピュータのメモリにダウンロードされてもよい。

コンピュータに内蔵されるCPUが文書検索プログラムを実行することによって、コンピュータは検索装置として機能する。

(実施の形態12)

5 実施の形態12では、検索キーワードとして指定された文字要素列に含まれる文字要素の数と、検索キーワードとして指定された文字要素列に含まれる文字要素のうち認識結果に一致した文字要素の数とに基づいて、検索結果が検索キーワードに一致する確率(評価値)が取得される。この確率(評価値)に基づいて、検索結果の正当性が判定される。

以下の説明では、「文字要素」を簡略化して「文字」という。

10 通常、さまざまな言語の単語には冗長性があり、数文字が分からない場合でも単語が特定できる場合が多い。この傾向は、単語を構成する文字数が多ければ多いほど当てはまる。本実施の形態は、単語のこのような傾向を利用して誤りを含んだ文字列から単語を検索できることを示す。

15 以下、図26を参照して、「・・・オックスフォード大学は・・・」というオリジナル文書を文字認識することによって得られた「・・・オッタスフォード大学は・・・」という認識結果から検索語「オックスフォード」を検索する検索処理を説明する。ここで、認識結果は、文書データとしてメモリ140に格納される。メモリ140は、任意のタイプの記憶媒体であり得る。

20 なお、認識結果の各文字には、認識結果の確からしさ(正解確率)を表す信頼度が付与されている。検索を行う前に、確率テーブルが予め作成される。

図27は、確率テーブルの一例を示す。確率テーブルは、パラメータ $n$ 、 $k$ に対して確率 $P_a(n, k)$ を取得するために使用される。

25 ここで、 $n$ は、検索語に含まれる文字の数を示し、 $k$ は、検索語に含まれる $n$ 文字のうち、検索対象の文書データ中の対応する文字と一致した文字の数を示す。確率 $P_a(n, k)$ は、検索結果が検索語に一致する確率を示す。

図27に示される確率テーブルは、誤りを含まない大量のテキストデータと単

語辞書とを用いて算出され得る。算出方法としては、まず、単語辞書中の単語  
( $n$ 文字)について、テキストデータ中のすべての連続する $n$ 文字に対して、単語  
と一致する文字数を調べ、一致する文字数別に累計 $N_k$  ( $i = 1, \dots, n$ )をとる。この累計 $N_k$ を用いると、 $n$ 文字の検索語のうち $k$ 文字が一致した  
5 場合に、その部分が検索語である確率 $P_a(n, k)$  ( $= N_k / N_n$ )が算出で  
きる。

この確率 $P_a(n, k)$ は、文字数 $n$ が一致していても単語の表記が異なる場  
合や、一致する文字数 $k$ の位置によって、当然異なるものとなるが、本実施の形  
態では、単語の表記や、一致する文字の位置に関しては区別をしていない。すな  
10 わち、文字数 $n$ が等しいすべての単語について、 $k$ 文字が一致する回数をそれぞ  
れ累計し、その累計の和(平均でもよい)を用いて算出している。

なお、単語の表記別や、一致した文字の位置別、単語を構成する字種別(漢字、  
ひらがな、カタカナ、英字など)にこのような確率を算出しておいて使用しても  
よい。

15 なお、単語の文字数が多い場合は、一致した文字数 $k$ がある程度になると、単語  
の文字数 $n$ に関わらず一致した文字数 $k$ だけで確率が表せる( $P_a(n, k)$   
において、 $n$ が変化してもほぼ一定)場合がある。この場合は、一致した文字数  
 $k$ のみに依存する確率 $P_a(k)$ を $P_a(n, k)$ の代わりに用いてもよい。

20 図26に示される例では、検索語「オックスフォード」に対し、検索対象の文  
書データにおいて誤認識の「夕」(信頼度0.42)を除くすべての文字が一致  
する。

ここで、この照合箇所(検索結果)の正当性を表す $P_w$ は、(式1)によって  
表される。

25 
$$P_w = P_a(n, k) \cdot P_b(k) \dots (\text{式1})$$

(式1)において、 $P_a(n, k)$ は、 $n$ 文字の単語のうち $k$ 文字が一致した場合に、その検索結果が検索語である確率である。この場合には、「オックスフォード」という $n=8$ 文字の単語のうち $k=7$ 文字が一致しているため、図27から $P_a(8, 7)=0.9$ となる。

5       また、 $P_b(k)$ は、 $k$ 文字がすべて検索対象の文字である確率を表す。各文字に付与されている信頼度は確率であるから、認識結果と一致した文字の信頼度の積とすればよい。図26から、 $P_b(7)=0.98 \times 0.97 \times 0.99 \times 0.98 \times 0.99 \times 0.97 \times 0.96=0.85$ となる。

よって、この照合箇所(検索結果)の正当性を表す値は、 $P_w=P_a(8, 7) \times P_b(7)=0.9 \times 0.85=0.765$ となる。 $P_w$ の値は、予め定められた閾値(本実施の形態では0.6とする)よりも大きい。従って、この照合箇所(検索結果)の正当性が肯定される。

15       なお、長い文字数を持った単語の場合は、信頼度の積 $P_b(k)$ が小さくなりやすいため、何らかの方法で文字数に対して正規化してもよい。また、各文字の信頼度が確率でない場合は、確率に変換して用いたり、単純に平均を求めて $P_b(k)$ の代わりにしてもよい。

20       また、図28に示されるように、正解した文字が7文字であっても低い信頼度を持った文字が存在する場合は、その文字をカウントしない場合が結果的に $P_w$ としては大きくなる場合があるため、 $P_w$ が最大となるような正解文字数 $k$ を選ぶようにしてもよい(7文字正解とすると、 $P_w=P_a(8, 7) \times P_b(7)=0.90 \times (0.98 \times 0.97 \times 0.99 \times 0.98 \times 0.99 \times 0.97 \times 0.30)=0.239$ だが、8文字目の正解を含めないで6文字正解とすると $P_w=P_a(8, 6) \times P_b(6)=0.85 \times (0.98 \times 0.97 \times 0.99 \times 0.98 \times 0.99 \times 0.97)=0.752$ であるため、後者を $P_w$ として採用する)。

25       なお、文字毎に信頼度が付与されていないデータベースの場合は、検索語の長

さ  $n$  と、一致した文字数  $k$  を用いて  $P_a(n, k)$  を、照合した箇所の正当性を表す値としてもよい。

なお、一致しなかった文字は今回情報として用いていないが、一致しなかった文字の信頼度が高い場合はその文字が正解である可能性が高い（すなわち、1文字だけ異なる単語が存在し、検索結果としては正当性が低い）と考えられるので、一致しない文字の信頼度を利用したペナルティを導入してもよい（一致しない文字の信頼度が予め定めた閾値よりも高い、または、そのような文字が予め定めた文字数よりも多い場合は、照合箇所の正当性を表す  $P_w$  が閾値より大きくても検索結果として採用しないなど）。

このように、単語の冗長性を利用して、すべての文字が一致しなくても誤認識を含んだテキスト文書中から検索が可能となる。また、一致しない文字の個数をどう決めたらよいかという問題も、実際の大量のテキストデータベースから図 27 のような確率テーブルを用いて（式 1）で検索箇所の正当性を数値化することによって解決することができる。

### （実施の形態 13）

実施の形態 13 では、実施の形態 12 で説明した確率テーブルに加えて、実施の形態 3 で説明した文字要素間の距離テーブルを用いて、検索処理が実行される。

以下の説明では、「文字要素」を簡略化して「文字」という。

以下、図 29 を参照して、「・・・オックスフォード大学は・・・」というオリジナル文書を文字認識することによって得られる「・・・オッタヌフォード大学は・・・」という認識結果から検索語「オックスフォード」を検索する検索処理を説明する。認識結果は、文書データとしてメモリ 140 に格納される。メモリ 140 は、任意のタイプの記憶媒体であり得る。

認識結果の各文字には、認識結果の確からしさ（正解確率）を表す信頼度が付与されている。実施の形態 12 と同様に、検索を行う前に図 27 に示される確率テーブルが算出される。

検索時には、実施の形態 1 と同様に、各文字に付与された信頼度に基づいて、文字要素間の距離と比較される基準距離が決定される。実施の形態 3 と同様に、基準距離に基づいて、発生頻度（確率）が決定される。

照合箇所（検索結果）の正当性を表す  $P_w$  は、（式 2）によって表される。

5

$$P_w = P_a(n, k) \quad \dots \quad (\text{式 2})$$

ここで、確率  $P_a(n, k)$  は、図 27 に示される確率テーブルにおいて定義されている。

10

認識結果そのものが一致した文字数は 1, 2, 5, 6, 7, 8 文字目の 6 文字であるが、3 文字目に関しては文字要素間テーブルを参照することにより、結果的に 7 文字が一致していることになる（信頼度 0.42 の認識結果「夕」は、「ク」である確率（頻度）が  $0.3 > 0$  である）。よって、 $P_w = P_a(8, 7) = 0.9$  となる。ここで、予め定められた閾値を 0.80 とすると、 $P_w$  の値はこの閾値より大きいため、検索結果の正当性を肯定するようにしてもよい。

15

しかし、検索の目的によっては検索したい文字列以外の文字列が検索されてしまう“検索ノイズ”をできるだけ減らしたいということもある。その場合は、より詳細な正当性の判定として、一致しなかった 4 文字目の「ヌ」について、距離の基準値（基準距離）を大きくする（すなわち、認識信頼度を小さく設定しなおし、距離の基準値（基準距離）を求める）ことにより、距離の基準値（基準距離）が 20 のときには、発生頻度が 0 となるために検出することができなかった認識誤りを検出することが可能になる。

20

一致しなかった文字に対して、すべての文字の可能性を許容したワイルドカードの扱いにすると、偶然 1 文字だけ異なる別の単語が検索されてしまう可能性が大きい。しかし、距離の基準値（基準距離）を少しだけ大きい値に再設定することにより、文字要素間のテーブルにおいて誤認識しやすい類似文字のみ、誤認識

25



を想定して文字の検索を行うことが可能になる。その結果、検索ノイズを減らすことができる。

また、距離の基準値（基準距離）を大きい値に再設定することによって、誤認識で認識信頼度が大きい場合に発生頻度が0となる場合（信頼度が大きいがため  
5 に距離の基準値（基準距離）が小さめになるが、距離の基準値（基準距離）がもう少し大きければ頻度 $>0$ となる場合）を救済することができる。

なお、検索結果の正当性を表すPwの値に応じて、距離の基準値（基準距離）を大きくする幅（認識信頼度の下げ幅）を制御してもよい。また、そのような距離の基準値を大きくする（認識信頼度を小さくする）文字数を制御してもよい。

10 なお、検索結果の正当性を表すPwの値に応じて、距離の基準値（基準距離）を大きくするか、ワイルドカード扱いにするかを制御してもよい。

図30は、実施の形態13のあいまい検索処理の手順を示す。このあいまい検索処理は、CPU110によって文書検索プログラムに従って実行される。

15 はじめに、文字要素間の距離と比較される基準距離（初期値）が設定される（ステップS3001）。この基準距離は、検索語に含まれる文字ごとに設定してもよいし、検索語に対して共通に設定してもよい。

1 文字単位の照合が行われ（ステップS3002）、その照合結果に基づいて、検索語に相当する認識結果文字列の評価値が計算される（ステップS3003）。例えば、評価値として（式1）および（式2）に示されるPwを使用することが  
20 できる。

次に、検索語に含まれるすべての文字が認識結果に一致しているか否かが判定される（ステップS3004）。

ステップS3004における判定結果が「Yes」の場合には、評価値と所定の閾値1とが比較される（ステップS3005）。

25 評価値が所定の閾値1より大きい場合には、認識結果から検索語が検出されたと判定される（ステップS3006）。

評価値が所定の閾値 1 以下である場合には、検索結果が棄却（リジェクト）される（ステップ S 3 0 0 7）。このように検索結果を棄却する理由は、認識結果に含まれる文字の中に低い信頼度が付与された文字が多い場合に発生しやすい誤検出を抑制するためである。

- 5       ステップ S 3 0 0 4 における判定結果が「No」の場合には、評価値と所定の閾値 2 とが比較される（ステップ S 3 0 0 8）。

10       評価値が所定の閾値 2 より大きい場合には、基準距離が変更され（ステップ S 3 0 0 9）、認識結果と一致しなかった検索語の文字について、1 文字単位の照合がやり直される（ステップ S 3 0 0 2）。ステップ S 3 0 0 9 において、基準距離の変更は、認識結果と一致しなかった検索語の文字についてのみ行われる。なお、基準距離の変更が検索語のすべての文字について行われてもよい。また、基準距離は初期値より大きい一定の値に変更されてもよいし、評価値に応じて変動する値（初期値より大きい可変値）に変更されてもよい。

- 15       評価値が所定の閾値 2 以下である場合には、認識結果から検索語が検出されなかったと判定される（ステップ S 3 0 1 0）。

20       なお、ステップ S 3 0 0 9 からステップ S 3 0 0 2 に回帰する回数の上限值  $n$  が予め設定される。本実施の形態では、 $n = 2$  である。上限値  $n$  を設ける理由は、評価値がいったん所定の閾値 2 を越えると、検索語に含まれるすべての文字が認識結果に一致するまで、ステップ S 3 0 0 9 において基準距離の変更が行われてステップ S 3 0 0 2 に回帰してしまうため、基準距離をいくら増やしても検索語に含まれるすべての文字が認識結果に一致しない場合には、無限ループに陥る可能性があるからである。上限値  $n$  を設けることにより、そのような無限ループに陥ることを回避することが可能になる。

- 25       なお、確からしい検索結果の表示方法と、そうでない検索結果の表示方法とを区別することにより、検索結果の確からしさを分かりやすく表示することが好ましい。例えば、評価値に応じて、検索結果の表示方法を変更するようにすればよ

い。

なお、実施の形態 1 2 で説明した検索処理は、図 3 0 に示されるステップ S 3 0 0 1 の処理と、ステップ S 3 0 0 8 の判定結果が「Y e s」の場合の処理を除いて、実施の形態 1 3 で説明した検索処理と同一である。実施の形態 1 2 では、

5 ステップ S 3 0 0 8 の判定結果が「Y e s」の場合、認識結果から検索語が検出されたと判定される。その後、検索処理は終了する。

(実施の形態 1 4)

実施の形態 1 4 は、実施の形態 1 2、1 3 で説明した検索処理の改良である。

以下の説明では、「文字要素」を簡略化して「文字」という。

10 以下、図 3 1 を参照して、「・・・オックスフォードの学生達・・・」というオリジナル文書を文字認識することによって得られる「・・・オッタスフォード○学生達・・・」という認識結果から検索語「オックスフォード大学」を検索する検索処理を説明する。認識結果は、文書データとしてメモリ 1 4 0 に格納される。メモリ 1 4 0 は、任意のタイプの記憶媒体であり得る。

15 認識結果の各文字には、認識結果の確からしさ（正解確率）を表す信頼度が付与されている。実施の形態 1 2、1 3 と同様に、検索を行う前に図 2 7 に示される確率テーブルが算出される。

検索時には、まず検索語「オックスフォード大学」が、複数の単語に分割され得るか否かが調べられる。この判定には、例えば、予め用意された単語辞書が使用される。本実施の形態では、予め用意された単語辞書には、「オックスフォード」「大学」という単語が存在していると仮定する。この場合、検索語「オックスフォード大学」は、「オックスフォード」+「大学」という 2 つの単語に分割される。

20

検索時には、「オックスフォード」という検索語の後に、「大学」という検索語がある場所が検索対象の文書データの中から探索される。各検索語についての探索のしかたは、実施の形態 1 2、1 3 で説明したものと同様である。

25

もし、2つの単語に分割しないで検索した場合、図31に示される例では、「オックスフォード大学」という10文字の単語において1, 2, 4, 5, 6, 7, 8, 10文字目の8文字が一致する。また、通常Pa(10, 8)は非常に大きいため、誤検出されて検索ノイズとなり易い。他の場所でも、単語「オックスフォード」が存在する近辺では検索がヒットし易くなり、検索ノイズが頻発してしまうという恐れがある。このように、複数の単語で長い検索語が形成されている場合には、検索語を、単語辞書などを用いて複数の単語に分割することにより、検索ノイズを低減することができる。

なお、単語辞書には、通常の単語だけでなく、複数の単語に共通して表れ易い部分文字列が含まれていることが好ましい。例えば、「プランテーション」「オリエンテーション」「ステーション」などに含まれる「テーション」という文字列を辞書に含んでおき、検索時には、それぞれの文字列を「プラン」+「テーション」、あるいは、「オリエン」+「テーション」、あるいは、「ス」+「テーション」と分割してそれぞれ検索することによって、検索語の一部が等しい別の単語が誤検出されるのを防ぐことができる。

#### 産業上の利用可能性

本発明によれば、文字要素と文字要素との間に、文字要素と文字要素との類似度に関連する距離が予め設定されている。この文字要素間の距離と所定の基準距離との比較結果に基づいて、認識結果に含まれる文字要素が検索キーワードに含まれる文字要素に一致するか否かが判定される。所定の基準距離を認識結果の信頼度に応じて可変にすることにより、認識結果に応じて許容可能な誤り度合を動的に変更しながら検索を行うことが可能になる。

また、文字要素間の距離をテーブルの形式で予め設定しておくことにより、検索時に複雑な距離計算を行う必要がない。その結果、検索を高速に行うことが可能にある。

- 5      また、特定の文字要素に対して、その特定の文字要素に接続される可能性のある複数の文字要素を決定することにより、オリジナル文書のレイアウトを誤って認識した場合でも、認識結果から検索キーワードを正しく検出することが可能になる。その結果、オリジナル文書の文章が縦書きであるか横書きであるかを誤って認識した場合や、改行後に継続する行を誤って認識した場合でも、認識結果から検索キーワードを正しく検出することが可能になる。

## 請求の範囲

1. 文字列を文字認識することによって得られる第1の文字要素列から第2の文字要素列を検索する検索方法であって、

5 前記第1の文字要素列は、第1の文字要素を含み、前記第2の文字要素列は、第2の文字要素を含み、

前記第1の文字要素と前記第2の文字要素との間には、前記第1の文字要素と前記第2の文字要素との類似度に関連する距離が予め設定されており、

前記検索方法は、

10 前記距離と所定の基準距離とを比較するステップと、

前記距離と前記所定の基準距離との比較結果に基づいて、前記第2の文字要素が前記第1の文字要素に一致するか否かを判定するステップと

を包含する、検索方法。

15 2. 前記第1の文字要素には文字認識の信頼度が予め設定されており、

前記所定の基準距離は、前記信頼度に基づいて決定される、請求項1に記載の検索方法。

20 3. 前記所定の基準距離は、ユーザからの入力に基づいて決定される、請求項1に記載の検索方法。

4. 前記検索方法は、

前記所定の基準距離を新たな基準距離に変更するステップと、

前記距離と前記新たな基準距離とを比較するステップと、

25 前記距離と前記新たな基準距離との比較結果に基づいて、前記第2の文字要素が前記第1の文字要素に一致するか否かを判定するステップと

をさらに包含する、請求項 1 に記載の検索方法。

5. 前記第 1 の文字要素と前記第 2 の文字要素との間には、前記第 1 の文字要素と前記第 2 の文字要素との類似度に関連する複数の距離が予め設定されており、

5 前記複数の距離のうち選択された 1 つが前記距離として使用される、請求項 1 に記載の検索方法。

6. 前記複数の距離のうちの 1 つは、ユーザからの入力に基づいて選択される、請求項 5 に記載の検索方法。

10

7. 前記距離は、確率的な分布を有している、請求項 1 に記載の検索方法。

8. 文字列を文字認識することによって得られる第 1 の文字要素列から第 2 の文字要素列を検索する検索方法であって、

15

前記第 1 の文字要素列は、複数の文字要素を含み、

前記複数の文字要素のうち特定の文字要素に対して、前記特定の文字要素に接続される可能性のある複数の文字要素が予め決定されており、

前記検索方法は、

20

前記複数の文字要素のうち特定の文字要素と、前記特定の文字要素と異なる前記複数の文字要素のうちの 1 つとを接続することによって得られる文字要素列が、前記第 2 の文字要素列の少なくとも一部に一致するか否かを判定するステップを包含する、検索方法。

9. 前記検索方法は、

25

前記特定の文字要素に接続される可能性のある前記複数の文字要素から 1 つの文字要素を選択するステップと、

前記特定の文字要素と前記選択された文字要素とを接続することによって得られる文字要素列が、前記第 2 の文字要素列の少なくとも一部に一致するか否かを判定するステップと

を包含する、請求項 8 に記載の検索方法。

5

10. 前記特定の文字要素は、行または列の末尾に配置されており、前記特定の文字要素に接続される可能性のある前記複数の文字要素のそれぞれは、行または列の先頭に配置されている、請求項 8 に記載の検索方法。

10

11. 前記特定の文字要素と、前記特定の文字要素に接続される可能性のある前記複数の文字要素のうちの 1 つとは、同一の行または列に配置されており、

前記特定の文字要素と、前記特定の文字要素に接続される可能性のある前記複数の文字要素のうちの他の 1 つとは、異なる行または列に配置されており、かつ、同一の列または行に配置されている、請求項 8 に記載の検索方法。

15

12. 文字列を文字認識することによって得られる第 1 の文字要素列から第 2 の文字要素列を検索する検索方法であって、

前記第 1 の文字要素列は、少なくとも 1 つの第 1 の文字要素を含み、前記第 2 の文字要素列は、少なくとも 1 つの第 2 の文字要素を含み、

20

前記検索方法は、

前記第 2 の文字要素列に含まれる前記第 2 の文字要素の数と、前記第 2 の文字要素列に含まれる前記第 2 の文字要素のうち、前記第 1 の文字要素のうち対応する第 1 の文字要素に一致した第 2 の文字要素の数とに基づいて、検索結果が前記第 2 の文字要素列に一致する確率を取得するステップと、

25

前記確率に基づいて、前記検索結果の正当性を判定するステップと  
を包含する、検索方法。



1 3. 前記第 2 の文字要素と前記対応する第 1 の文字要素との間には、前記第 2 の文字要素と前記対応する第 1 の文字要素との類似度に関連する距離が予め設定されており、

5 前記検索方法は、

前記距離と所定の基準距離とを比較するステップと、

前記距離と前記所定の基準距離との比較結果に基づいて、前記第 2 の文字要素が前記対応する第 1 の文字要素に一致するか否かを判定するステップと

をさらに包含する、請求項 1 2 に記載の検索方法。

10

1 4. 前記検索方法は、

前記第 2 の文字要素列に含まれる前記第 2 の文字要素のうち、前記第 1 の文字要素のうち対応する第 1 の文字要素に一致しなかった第 2 の文字要素について、所定の基準距離を再設定した後に、再設定された所定の基準距離を用いて、前記第 2 の文字要素が前記対応する第 1 の文字要素に一致するか否かを再判定するステップ

15

をさらに包含する、請求項 1 3 に記載の検索方法。

1 5. 前記検索方法は、

20

前記第 2 の文字要素列を複数の部分に分割するステップ

をさらに包含する、請求項 1 2 に記載の検索方法。

1 6. 文字列を文字認識することによって得られる第 1 の文字要素列から第 2 の文字要素列を検索する検索装置であって、

25

前記第 1 の文字要素列は、第 1 の文字要素を含み、前記第 2 の文字要素列は、第 2 の文字要素を含み、

前記第 1 の文字要素と前記第 2 の文字要素との間には、前記第 1 の文字要素と前記第 2 の文字要素との類似度に関連する距離が予め設定されており、

前記検索装置は、

前記距離と所定の基準距離とを比較する手段と、

- 5 前記距離と前記所定の基準距離との比較結果に基づいて、前記第 2 の文字要素が前記第 1 の文字要素に一致するか否かを判定する手段と  
を備えている、検索装置。

- 10 17. 文字列を文字認識することによって得られる第 1 の文字要素列から第 2 の文字要素列を検索する検索装置であって、

前記第 1 の文字要素列は、複数の文字要素を含み、

前記複数の文字要素のうち特定の文字要素に対して、前記特定の文字要素に接続される可能性のある複数の文字要素が予め決定されており、

前記検索装置は、

- 15 前記複数の文字要素のうち特定の文字要素と、前記特定の文字要素と異なる前記複数の文字要素のうちの 1 つとを接続することによって得られる文字要素列が、前記第 2 の文字要素列の少なくとも一部に一致するか否かを判定する手段  
を備えた、検索装置。

- 20 18. 文字列を文字認識することによって得られる第 1 の文字要素列から第 2 の文字要素列を検索する検索装置であって、

前記第 1 の文字要素列は、少なくとも 1 つの第 1 の文字要素を含み、前記第 2 の文字要素列は、少なくとも 1 つの第 2 の文字要素を含み、

前記検索装置は、

- 25 前記第 2 の文字要素列に含まれる前記第 2 の文字要素の数と、前記第 2 の文字要素列に含まれる前記第 2 の文字要素のうち、前記第 1 の文字要素のうち対応す

る第 1 の文字要素に一致した第 2 の文字要素の数とに基づいて、検索結果が前記第 2 の文字要素列に一致する確率を取得する手段と、

前記確率に基づいて、前記検索結果の正当性を判定する手段と  
を備えた、検索装置。

5

19. 文字列を文字認識することによって得られる第 1 の文字要素列から第 2 の文字要素列を検索する検索処理をコンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、

10 前記第 1 の文字要素列は、第 1 の文字要素を含み、前記第 2 の文字要素列は、第 2 の文字要素を含み、

前記第 1 の文字要素と前記第 2 の文字要素との間には、前記第 1 の文字要素と前記第 2 の文字要素との類似度に関連する距離が予め設定されており、

前記検索処理は、

前記距離と所定の基準距離とを比較するステップと、

15 前記距離と前記所定の基準距離との比較結果に基づいて、前記第 2 の文字要素が前記第 1 の文字要素に一致するか否かを判定するステップと  
を包含する、記録媒体。

20 20. 文字列を文字認識することによって得られる第 1 の文字要素列から第 2 の文字要素列を検索する検索処理をコンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、

前記第 1 の文字要素列は、複数の文字要素を含み、

前記複数の文字要素のうち特定の文字要素に対して、前記特定の文字要素に接続される可能性のある複数の文字要素が予め決定されており、

25 前記検索処理は、

前記複数の文字要素のうち特定の文字要素と、前記特定の文字要素と異なる前

記複数の文字要素のうちの1つとを接続することによって得られる文字要素列が、前記第2の文字要素列の少なくとも一部に一致するか否かを判定するステップを包含する、記録媒体。

- 5        21. 文字列を文字認識することによって得られる第1の文字要素列から第2の文字要素列を検索する検索処理をコンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、

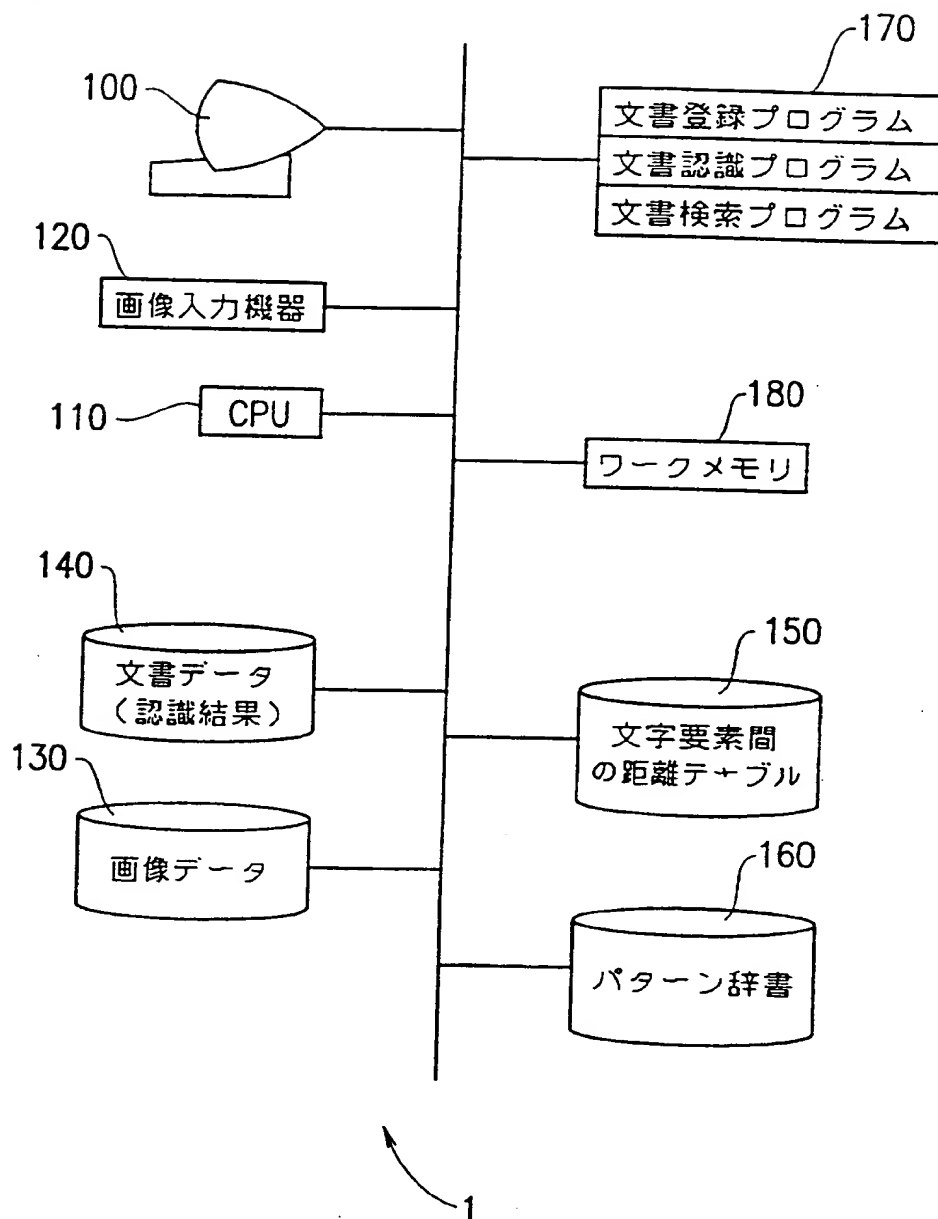
前記第1の文字要素列は、少なくとも1つの第1の文字要素を含み、前記第2の文字要素列は、少なくとも1つの第2の文字要素を含み、

- 10        前記検索処理は、

前記第2の文字要素列に含まれる前記第2の文字要素の数と、前記第2の文字要素列に含まれる前記第2の文字要素のうち、前記第1の文字要素のうち対応する第1の文字要素に一致した第2の文字要素の数とに基づいて、検索結果が前記第2の文字要素列に一致する確率を取得するステップと、

- 15        前記確率に基づいて、前記検索結果の正当性を判定するステップとを包含する、記録媒体。

図 1



**THIS PAGE BLANK (USPTO)**

図 2

	亜	啞	𠂇	𠂈	00
亜		10	132	166	172
啞			115	152	164
𠂇				143	191
𠂈					69
00					

図 3A

「𠂇」, 「𠂈」

図 3B

「)𠂇」, 「𠂈1」

図 4

オリジナル文書	.... 日本的人口構成は ....
文字認識結果	.... 日木の人区構成は ....

**THIS PAGE BLANK (USPTO)**



図 5

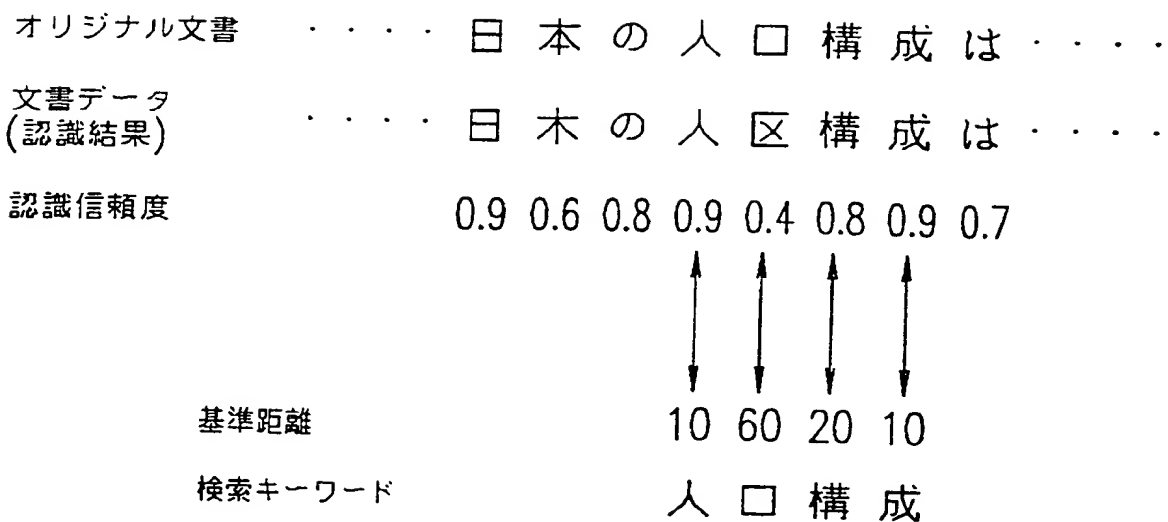


図 6

	人	口	構	成	区	同
人	0	170	250	210	99	113
口	170	0	244	168	50	100
構	250	244	0	142	198	184
成	210	168	142	0	137	152

**THIS PAGE BLANK (USPTO)**

図 7

	亜	啞	≡	冂	00
亜		12	130	170	168
啞			114	150	170
≡				147	190
冂					60
00					

図 8A

「木」, 「木」→「林」

「0」, 「0」→「∞」

図 8B

「川」→「1」, 「1」, 「1」

「い」→「し」, 「1」

**THIS PAGE BLANK (USPTO)**

図 9

	林	$\infty$	し 1	1 1 1	川
木木	10	221	190	156	152
川	155	165	91	9	
い	201	119	13	89	95
冂	149	188	98	133	137
00	215	12	105	169	172

**THIS PAGE BLANK (USPTO)**

図 10A

		T		
下	距離	10	20	30
	発生頻度 (確率)	0.2	0.6	0.2

図 10B

		T	F	ト	Γ	木
下	距離	20	60	90	125	130
	分散	10	10	30	25	20

図 10C

		T	F	ト	Γ	木
下	最短距離	10	50	63	102	110
	最大距離	30	70	122	151	165

**THIS PAGE BLANK (USPTO)**

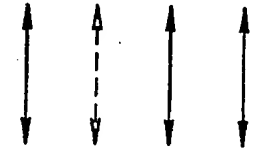


図 11

オリジナル文書	.....	日	本	の	人	口	構	成	は	.....
文書データ (認識結果)	.....	日	本	の	人	区	構	成	は	.....
認識信頼度		0.9	0.6	0.8	0.9	0.4	0.8	0.9	0.7	
					↑	↑	↑	↑		
					↓	↓	↓	↓		
基準距離					10	60	20	10		
発生頻度					0.9	0.1	0.9	0.9		
検索キーワード					人	口	構	成		

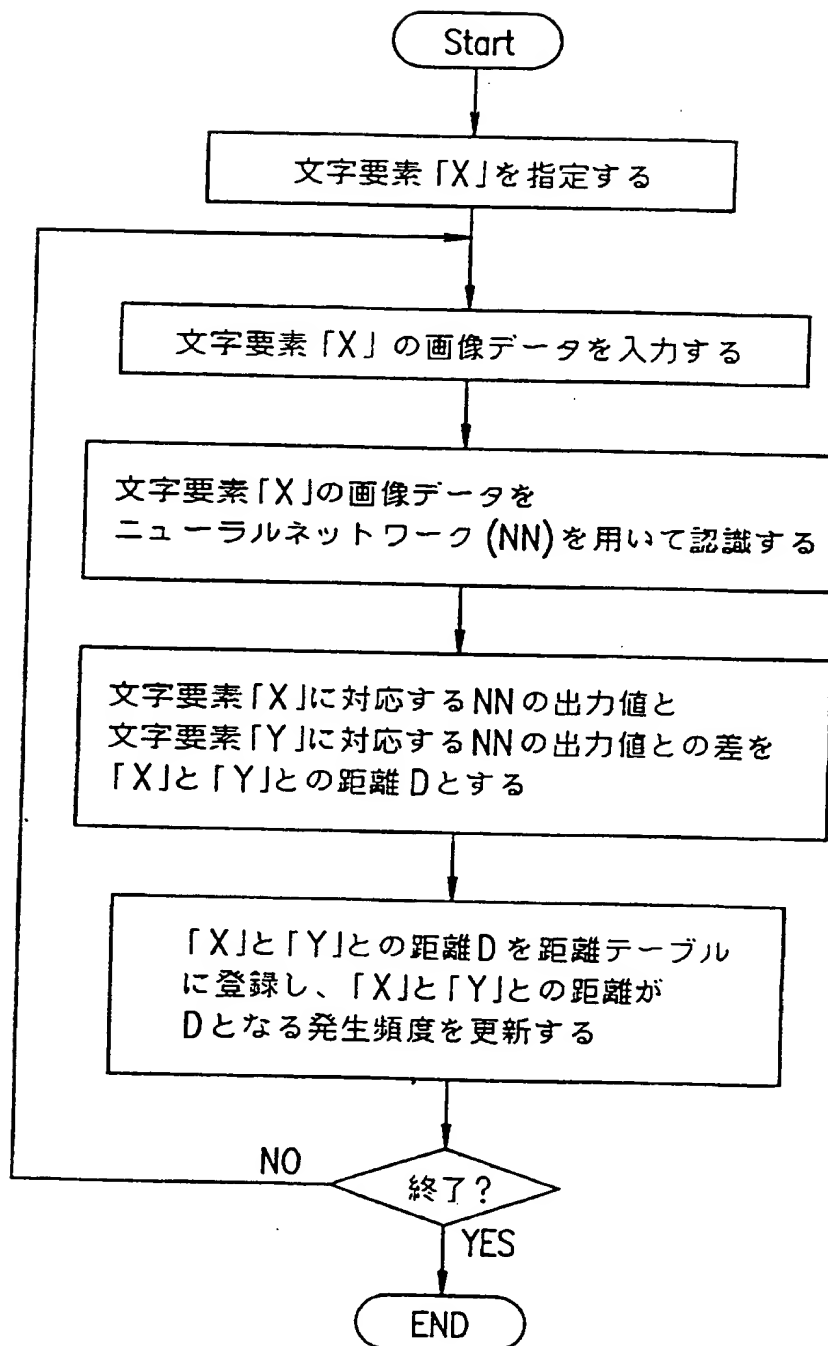
**THIS PAGE BLANK (USPTO)**

図 12

オリジナル文書	..... 日 本 の 人 口 構 成 は .....
文書データ (認識結果)	..... 日 本 の 人 口 構 成 は .....
認識信頼度	0.9 0.6 0.8 0.9 0.3 0.8 0.9 0.7
	
基準距離	10 80 20 10
発生頻度	0.9 0.0 0.9 0.9
検索キーワード	人 口 構 成

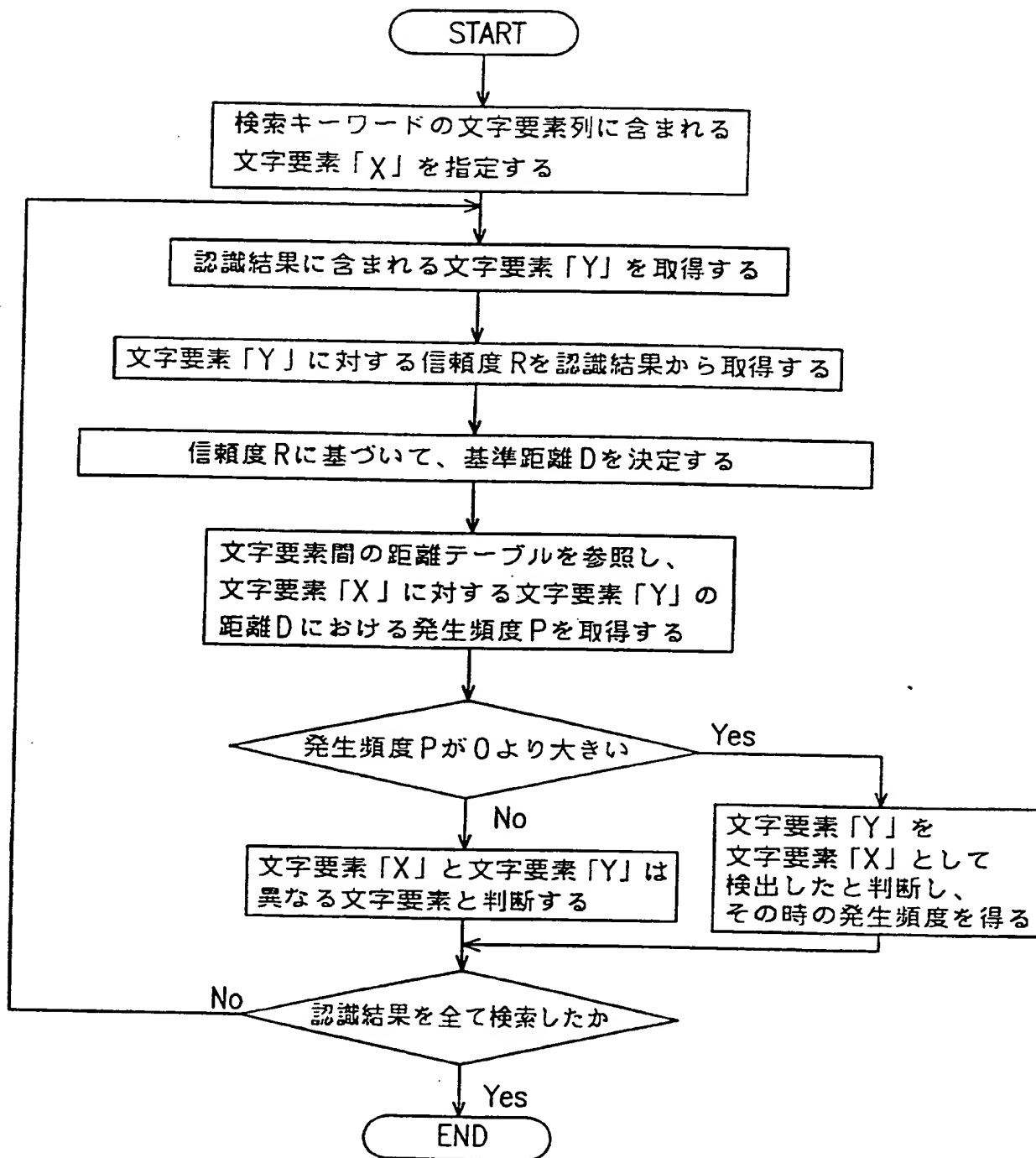
**THIS PAGE BLANK (USPTO)**

図 13



**THIS PAGE BLANK (USPTO)**

図 14



**THIS PAGE BLANK (USPTO)**



図 15

認識結果	.....	日	木	の	人	区	構	成	は	.....
候補		目	本	◎	入	凶	講	茂	ほ	
候補			大		ル	凶		感	ぼ	
候補						□				

**THIS PAGE BLANK (USPTO)**

**THIS PAGE BLANK (USPTO)**

図 16

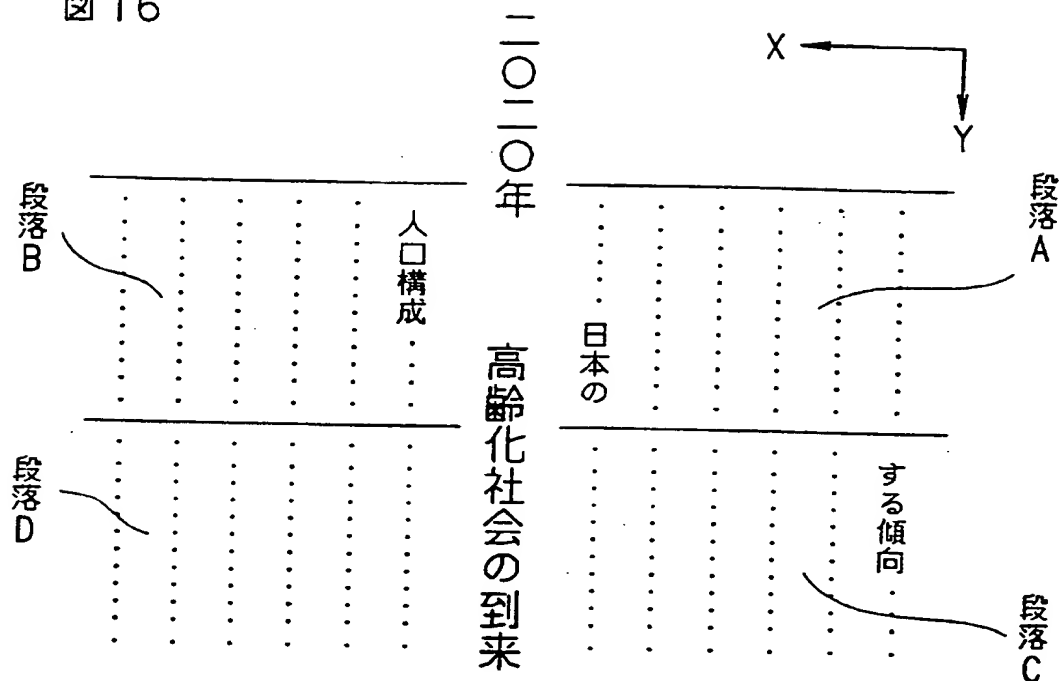
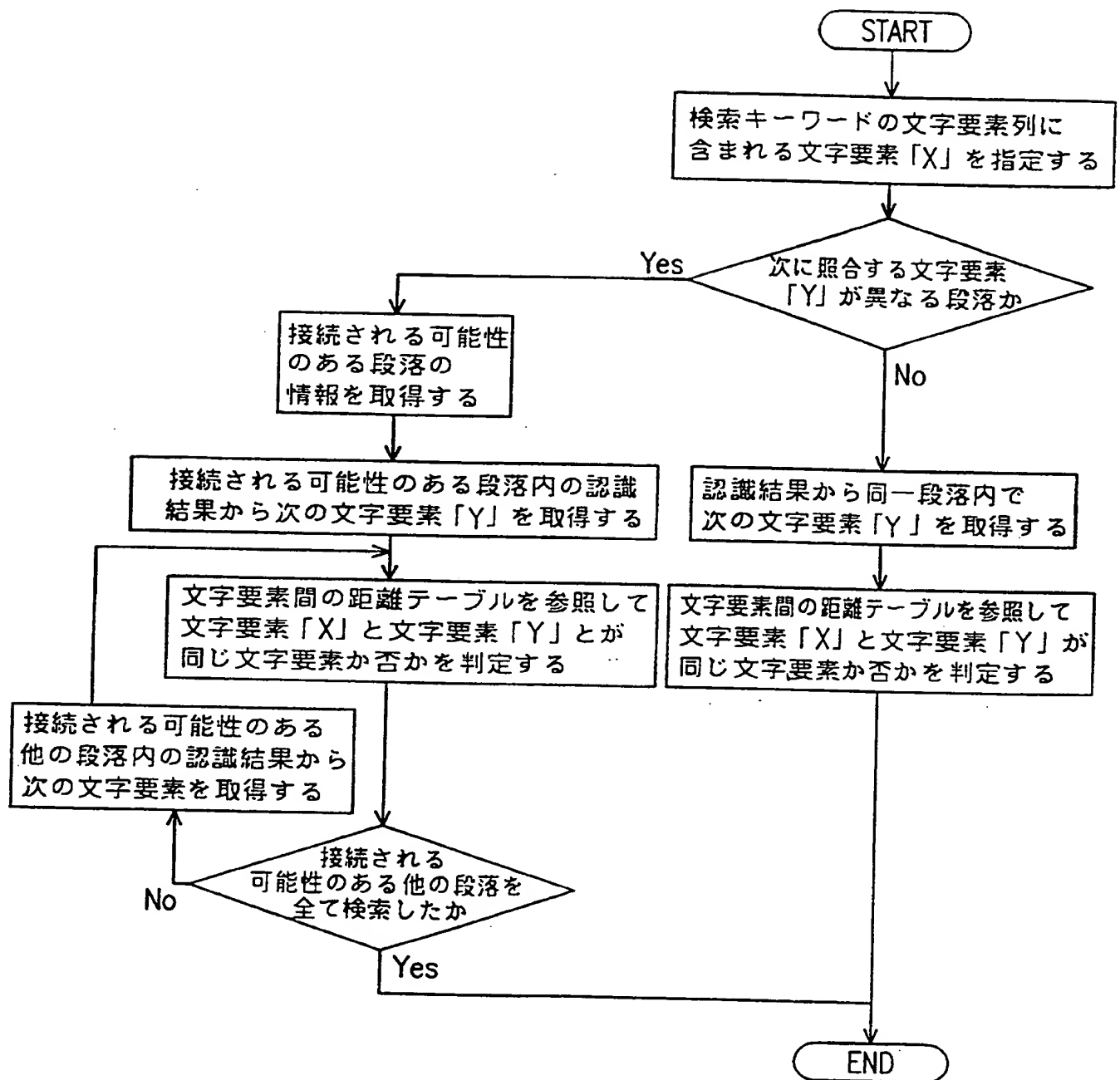


図 17

段落番号	接続される可能性のある段落の番号	段落単位の認識結果
A	B,C	..... 日本の
B	C,D	人口構成は.....
C	D	する傾向.....
D		.....

**THIS PAGE BLANK (USPTO)**

図 18



**THIS PAGE BLANK (USPTO)**

図 19

段落番号	段落単位の認識結果	段落の位置	
		x	y
A	..... 日本	10	100
B	人口構成は...	100	100
C	する傾向.....	10	200
D	.....	100	200

図 20

文字認識結果 1	...日本のする傾向...
文字認識結果 2	...日本の人口構成は...

図 21

京 都	29℃
大 阪	32℃
神 戸	30℃

**THIS PAGE BLANK (USPTO)**



図 22A

列番号	列単位の認識結果
1	℃℃℃
2	9 2 0
3	2 3 3
4	都 阪 戸
5	京 大 神

図 22B

行番号	行単位の認識結果
1	京都 29℃
2	大阪 32℃
3	神戸 30℃

図 23

オリジナル文書	．．．日本の人口構成は ．．．
文字認識結果	．．．日木の人口構成は ．．．

**THIS PAGE BLANK (USPTO)**

図 24

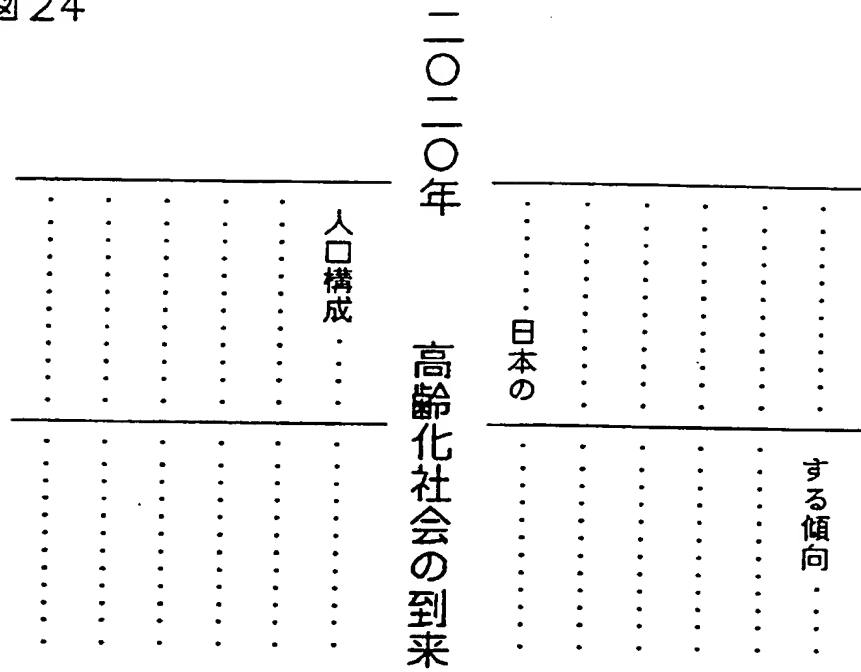


図 25

オリジナル文書	..... 日本の人口構成は .....
文字認識結果	..... 日本のする傾向 .....

**THIS PAGE BLANK (USPTO)**

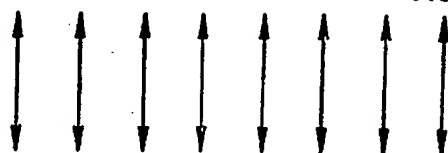
図 26

オリジナル文書 . . . オ ッ ク ス フ ォ ー ド 大 学 は . . .

文書データ  
(認識結果) . . . オ ッ タ ス フ ォ ー ド 大 学 は . . .

認識信頼度

0.98 0.42 0.98 0.97  
0.97 0.99 0.99 0.96



検索語

オ ッ ク ス フ ォ ー ド

図 27

$P_a(n, k)$

一致文字数 k \ 単語の文字数 n	1	2	3	4	5	6	7	8	...
1	1.0	0.1	0.1	0.1	0.08	0.05	0.03	0.01	
2		1.0	0.4	0.2	0.15	0.1	0.05	0.02	
3			1.0	0.6	0.4	0.3	0.2	0.1	
4				1.0	0.8	0.7	0.4	0.4	
5					1.0	0.9	0.8	0.8	
6						1.0	0.9	0.85	
7							1.0	0.9	
8								1.0	

**THIS PAGE BLANK (USPTO)**

図 28

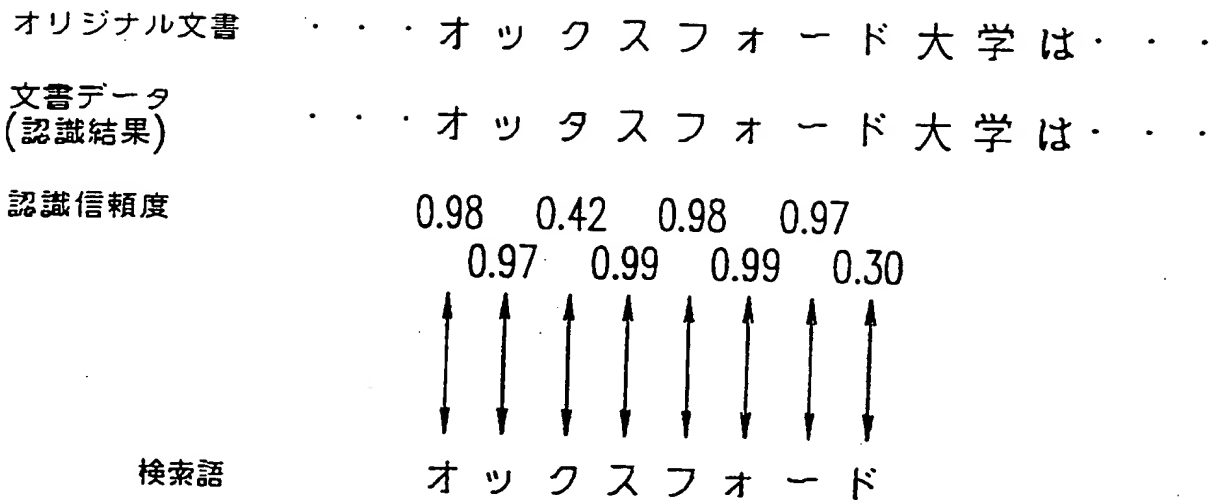
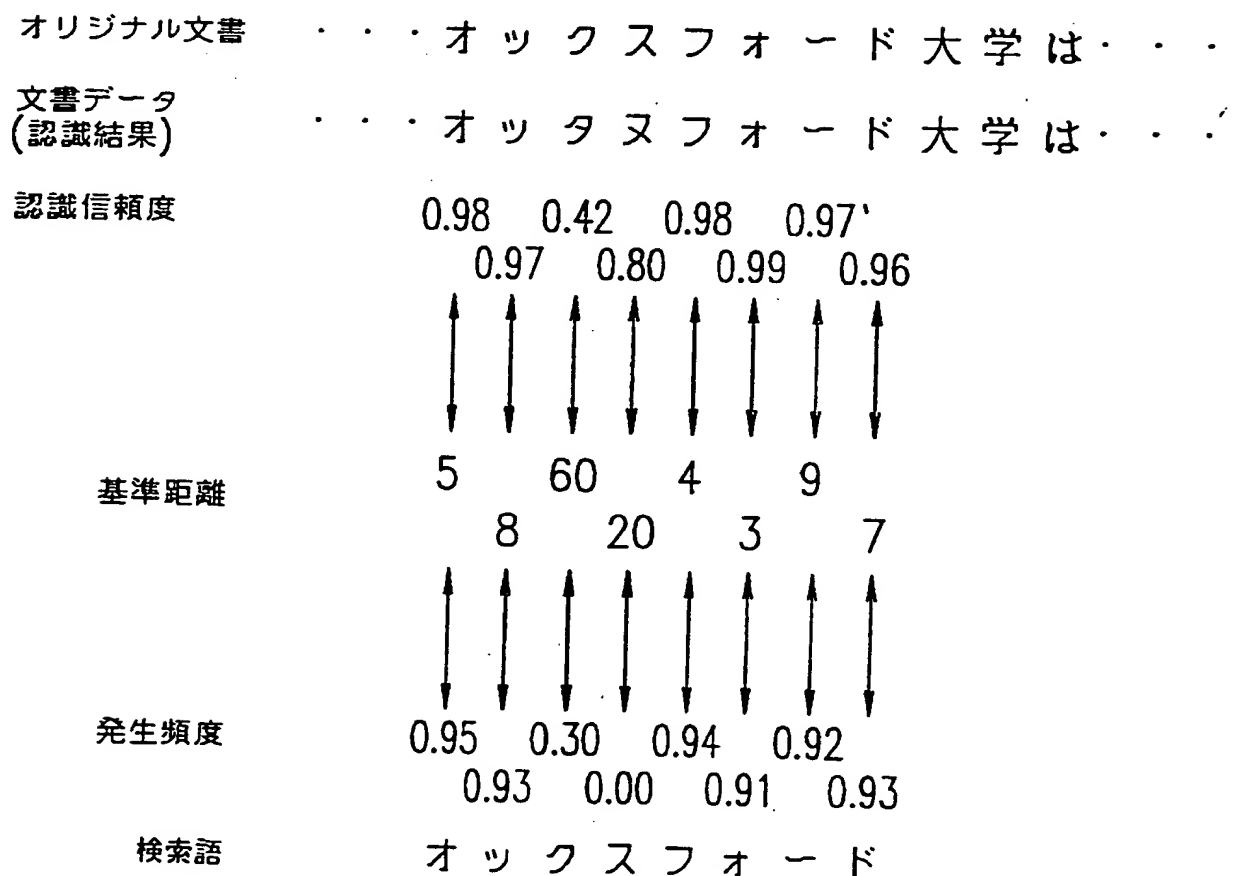


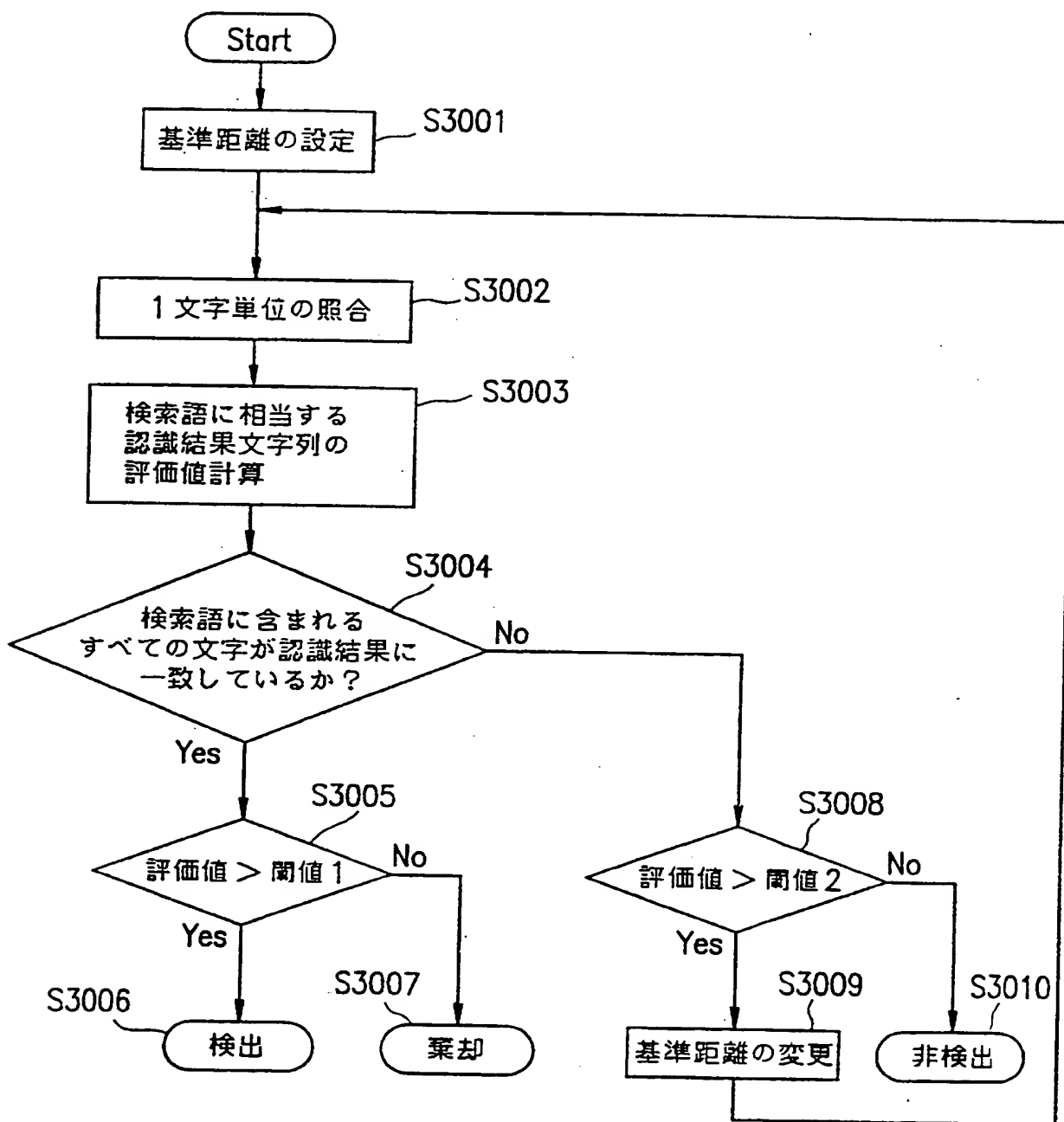
図 29



**THIS PAGE BLANK (USPTO)**



図 30



**THIS PAGE BLANK (USPTO)**

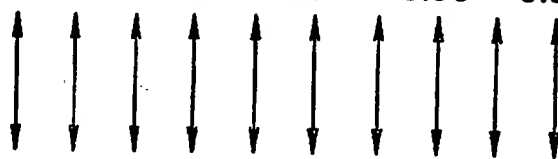
図 31

オリジナル文書 …… オックスフォードの学生達 ……

文書データ  
(認識結果) …… オッタスフォードの学生達 ……

認識信頼度

0.98 0.42 0.98 0.97 0.35  
0.97 0.99 0.99 0.98 0.98



検索語

オックスフォード大学

**THIS PAGE BLANK (USPTO)**

AMENDMENTS

(Amendment under Section 11 of the Japanese Law Concerning International Applications, Etc. Pursuant to the Patent Cooperation Treaty)

To the Commissioner of the Japanese Patent Office

1. Identification of the International Application  
PCT/JP99/07050

2. Applicant

Name MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD.  
Address 1006, Oaza Kadoma,  
Kadoma-shi, Osaka 571-8501 Japan  
Country of nationality Japan  
Country of residence Japan

3. Agent

Name (7828) Shusaku YAMAMOTO  
Address Fifteenth Floor, Crystal Tower, 2-27,  
Shiromi 1-chome, Chuo-ku,  
Osaka-shi, Osaka 540-6015 Japan

4. Item to be Amended  
Claims

5. Content of Amendments

As described in the annexed document,

(1) "a plurality of character elements" is amended to "character elements at a plurality of locations", and "the plurality of character elements" is amended to "the character elements at a plurality of locations" (lines 5 and 7 in claim 8) [corresponding to lines 8 and 14 in claim 8 of the English version],

**THIS PAGE BLANK (USPTO)**

(2) "a plurality of character elements" is amended to "character elements at a plurality of locations", and "the plurality of character elements" is amended to "the character elements at a plurality of locations" (lines 5 and 7 in claim 17) [corresponding to lines 8 and 14 in claim 17 of the English version],

(3) "a plurality of character elements" is amended to "character elements at a plurality of locations", and "the plurality of character elements" is amended to "the character elements at a plurality of locations" (lines 6 and 8 in claim 20) [corresponding to lines 9 and 15 in claim 20 of the English version],

6. List of Attached Documents

New sheets for "CLAIMS" on pages 51, 54, 55, and 56 [corresponding to pages 78, 79, 82, 83, 84, and 85 of the English language specification]

one copy of each

**THIS PAGE BLANK (USPTO)**



changing the first predetermined reference distance to a second reference distance;

comparing the distance with the second reference distance; and

5 determining whether the second character element matches the first character element based on a result of the comparison of the distance with the second reference distance.

10 5. A retrieval method according to claim 1, wherein a plurality of distances relevant to the similarity between the first character element and the second character element are predetermined between the first character element and the second character element, and

15 one distance selected from the plurality of distances is used as the distance.

20 6. A retrieval method according to claim 5, wherein the one of the plurality of distances is determined based on user input.

7. A retrieval method according to claim 1, wherein the distance has a probabilistic distribution.

25 8. A retrieval method for searching a first character element string obtained by subjecting a character string to character recognition for a second character element string,

30 wherein the first character element string includes a plurality of character elements,

for a specific character element of the plurality of character elements, a plurality of character elements having the possibility of being concatenated with the

**THIS PAGE BLANK (USPTO)**

specific character element are predetermined,

the retrieval method comprising the steps of:

determining whether a character element string  
obtained by concatenating the specific character element  
of the plurality of character elements with one character  
element of the plurality of character elements, the one  
character element being different from the specific  
character element, matches at least a part of the second  
character element string.

10

9. A retrieval method according to claim 8, comprising the  
steps of:

selecting one character element from the plurality  
of character elements having the possibility of being  
concatenated with the specific character element; and

15

determining whether a character element string  
obtained by concatenating the specific character element  
with the selected character element matches at least a part  
of the second character element string.

20

10. A retrieval method according to claim 8, wherein the  
specific character element is located at an end of a row  
or column, the plurality of character elements having the  
possibility of being concatenated with the specific  
character element are each located at a head of a row or  
column.

25

11. A retrieval method according to claim 8, wherein the  
specific character element and one of the plurality of  
character elements having the possibility of being  
concatenated with the specific character element are  
located at the same row or column, and

30

the specific character element and another one of

**THIS PAGE BLANK (USPTO)**

the retrieval device comprising:

means for comparing the distance with a predetermined reference distance; and

5 means for determining whether the second character element matches the first character element based on a result of the comparison of the distance with the predetermined reference distance.

10 17. A retrieval device for searching a first character element string obtained by subjecting a character string to character recognition for a second character element string,

wherein the first character element string includes a plurality of character elements, and

15 for a specific character element of the plurality of character elements, a plurality of character elements having the possibility of being concatenated with the specific character element are predetermined,

the retrieval device comprising:

20 means for determining whether a character element string obtained by concatenating the specific character element of the plurality of character elements with one character element of the plurality of character elements, the one character element being different from the specific  
25 character element, matches at least a part of the second character element string.

30 18. A retrieval device for searching a first character element string obtained by subjecting a character string to character recognition for a second character element string,

wherein the first character element string includes at least one first character element and the second character

**THIS PAGE BLANK (USPTO)**

element string includes at least one second character element,

the retrieval device comprising:

5 means for obtaining a probability that a search result matches the second character element string, based on the number of the second character elements included in the second character element string, and a number of the second character elements matching the corresponding first character elements out of the second character elements included in the second character element string; and  
10 means for determining the correctness of the search result based on the probability.

19. A computer readable recording medium in which a program  
15 for causing a computer to execute a retrieval process for searching a first character element string obtained by subjecting a character string to character recognition for a second character element string is recorded, and

wherein the first character element string includes  
20 a first character element and the second character element string includes a second character element,

a distance relevant to a similarity between the first character element and the second character element is predetermined between the first character element and  
25 the second character element,

the retrieval process comprising the steps of:  
comparing the distance with a predetermined reference distance; and

determining whether the second character element  
30 matches the first character element based on a result of the comparison of the distance with the predetermined reference distance.

**THIS PAGE BLANK (USPTO)**



20. A computer readable recording medium in which a program for causing a computer to execute a retrieval process for searching a first character element string obtained by  
5     subjecting a character string to character recognition for a second character element string is recorded,

          wherein the first character element string includes a plurality of character elements, and

          for a specific character element of the plurality of character elements, a plurality of character elements  
10     having the possibility of being concatenated with the specific character element are predetermined,

          the retrieval process comprising the steps of:

          determining whether a character element string obtained by concatenating the specific character element  
15     of the plurality of character elements with one character element of the plurality of character elements, the one character element being different from the specific character element, matches at least a part of the second character element string.

20

21. A computer readable recording medium in which a program for causing a computer to execute a retrieval process for searching a first character element string obtained by  
25     subjecting a character string to character recognition for a second character element string is recorded,

          wherein the first character element string includes at least one first character element and the second character element string includes at least one second character element,

30

          the retrieval process comprising the steps of:

          obtaining a probability that a search result matches the second character element string, based on the number of the second character elements included in the second

**THIS PAGE BLANK (USPTO)**

character element string, and a number of the second character elements matching the corresponding first character elements out of the second character elements included in the second character element string; and

- 5           determining the correctness of the search result based on the probability.

**THIS PAGE BLANK (USPTO)**

## REPLY TO WRITTEN OPINION

To Examiner of the Japanese Patent Office

1. Identification of the International Application  
PCT/JP99/07050

2. Applicant

Name                   MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD.  
Address               1006, Oaza Kadoma,  
                          Kadoma-shi, Osaka 571-8501 Japan  
Country of nationality   Japan  
Country of residence     Japan

3. Agent

Name                   (7828) Shusaku YAMAMOTO  
Address               Fifteenth Floor, Crystal Tower, 2-27,  
                          Shiromi 1-chome, Chuo-ku,  
                          Osaka-shi, Osaka 540-6015 Japan

4. Mailing Date        January 8, 2000

5. Content of Arguments

(1) Regarding "claims 1, 16, and 19 lack novelty in view of References 1 and 2"

In References 1 and 2, a probability that a specific character is recognized as the same character or other characters is held in a table (see Reference 1, Figure 1, page 627, and Reference 2, Figure 2, page 65), a score is calculated with reference to the probability in the table, and the score is compared with a threshold (see Reference 1, page 626, left paragraph, line 28).

In the present invention, the distance defined between character elements is obtained based on the shapes of character elements. Therefore, the distance is different from a probability (the present specification, page 13,

**THIS PAGE BLANK (USPTO)**

line 4) [corresponding to page 20, line 17 of the English version].

It is difficult to obtain the probability of an event which will unlikely occur based on the probability distributions in References 1 and 2. Moreover, the value of such a probability varies to a large extent depending on documents used for statistical samples. The distance between character elements of the present invention is easy to calculate by comparison of the shapes of the character elements (e.g., an input-output relationship of a character recognition system or a positional relationship in a feature quantity space (the present specification, page 13, line 4)) [corresponding to page 20, line 17 in the English version]. Therefore, a large number of statistical samples are not required, and a relationship between character elements can be obtained independent of such statistical samples.

There is no description regarding a distance between character elements in References 1 and 2. Therefore, the Applicant argues that the present invention has novelty and inventive step.

(2) Regarding "claim 2 lacks novelty in view of References 1 and 2"

In References 1 and 2, a probability that a specific character is recognized as the same character or other characters is held in a table (see Reference 1, Figure 1, page 627, and Reference 2, Figure 2, page 65), and a plurality of candidates having a high probability are produced with reference to a posterior probability that a character as a result of a recognition result is the same character or other characters from the table (Reference 1, page 626, left paragraph, line 28). As a result, the same set of candidates are produced for the same recognition result.

**THIS PAGE BLANK (USPTO)**



In the present invention, apart from the distance between character elements defined in advance, a reliability of a recognition result of each character element is provided. A reference distance is determined using the reliability. If the same recognition result has a different reliability, the tolerable level to recognition error is changed. As a result, a desirable search can be performed for an incorrect recognition result, and detection of extra character element strings can be suppressed (the present specification, page 16, line 26 to page 17, line 11) [corresponding to page 26, line 27 to page 27, line 4 of the English version].

In References 1 and 2, there is no description of the reliability given to a recognition result of an individual character element. Therefore, the present invention has novelty and inventive step.

(3) Regarding claims 3 through 7

Claims 3 through 7 are dependent from claim 1 and therefore have novelty and inventive step.

(4) Regarding "claims 8, 17, and 20 lack novelty in view of Reference 4"

Claims 8, 17, and 20 are amended in the Amendments submitted on the same date as this Reply.

In the Amendments, the phrase "a plurality of character elements having the possibility of being concatenated with a specific character element" is amended to the phrase "character elements at a plurality of locations having the possibility of being concatenated with a specific character element" in order to emphasize what was originally intended by the inventors.

**THIS PAGE BLANK (USPTO)**

Such an amendment clarifies that character elements existing at a plurality of (different) candidate locations (not fixed) which follow a specific character element (location). In conventional methods, as indicated in Reference 4, "a plurality of candidate characters for a recognition result of a character following a specific character element" are used.

There is no description regarding a retrieval method using "character elements at a plurality of locations having the possibility of being concatenated with a specific character element" in Reference 4. Therefore, the present invention has novelty and inventive step.

The retrieval method using "character elements at a plurality of locations having the possibility of being concatenated with a specific character element" is described in the specification of the present application, e.g., at page 30, lines 2-8 [corresponding to page 47, lines 5-18 of the English version] (an ending character element and leading character elements of a plurality of other paragraphs), at page 34, lines 8-15 [corresponding to page 53, line 24 to page 54, line 6 of the English version] (Figure 20; a recognition result is produced using character elements at a plurality of concatenating locations), and at page 35, lines 13-18 [corresponding to page 55, line 29 to page 56, line 8 of the English version] (Figure 22A and 22B; a recognition result is produced under an assumption of vertical and horizontal concatenation). As described above, a search is performed using character elements at a plurality of locations having the possibility of being concatenated with a specific character element. Therefore, even when concatenation is incorrectly determined in character recognition, a desirable result can be obtained in a character element string search.

**THIS PAGE BLANK (USPTO)**

(5) Regarding claim 9

Claim 9 is dependent from claim 8, and therefore has novelty.

(6) Regarding claims 10 and 11

Claims 10 and 11 are dependent from claim 8, and therefore have inventive step.

Search error is often caused by recognition error in character recognition. Therefore, it is a conventional objective to realize desirable search using a recognition result.

In general, a concatenation relationship between character elements is determined using a recognition result, independent of character recognition.

The present invention is significantly different from a conventional method for obtaining accurate character concatenation. The inventors focused their attention to a novel challenge to achieve a desired search, and made an attempt to improve search precision by a new approach in which candidates for a concatenation relationship, which have not been utilized prior to the present invention, are used after character recognition.

References 5 and 6 focus attention on that a correct character is selected from a plurality of candidates to be concatenated as accurately as possible. In this case, after the selection, information about the candidates cannot be utilized. In the present invention, such information about candidates which is not used after character recognition is newly utilized, thereby obtaining a desirable result in searching even if error occurs in concatenation.

**THIS PAGE BLANK (USPTO)**

(7) Regarding "claims 12, 13, 18, and 21 lack novelty in view of Reference 1"

In Reference 1, "a probability that a character string obtained by searching a text as a result of recognition is correct" is defined as

(a) a product of probabilities that each character in a character string is correct, or if matches do not occur in all characters,

(b) a partial certainty obtained by comparing between the certainty of each character element in a character string and multiplying the character elements except for a character element having the smallest certainty.

The method (b) in Reference 1 can be used only when a character is a wildcard, and the certainties of character elements except for a character element having the smallest certainty are only multiplied irrespective of the number of character elements in a character string to be matched. Ideally, the certainty of a character string as a whole should be higher when matches occur in nine out of ten characters in a character string (one character wildcard) than when a match occurs in one out of two characters in a character string (one character wildcard). This point is ignored in Reference 1. The product of certainties of characters excluding a character having the smallest certainty is only obtained in Reference 1.

In the inventions of claims 12, 13, 18, and 21, a novel probability is introduced based on the number of characters (character elements) in a search character string and the number of matched characters (character elements). Specifically, when a match occurs in one of the two characters in a character string, a probability that a character string as a result of recognition is a search character string is low, while when matches occur in nine

**THIS PAGE BLANK (USPTO)**



out of the ten characters in a character string, a probability that a character string as a result of recognition is a search character string is high. This fact is utilized in the present invention. Moreover, the number of characters used as wildcards is not limited to one.

Therefore, the Applicant argues that claims 12, 13, 18, and 21 have novelty over Reference 1.

(8) Regarding "claim 14 lacks inventive step in view of References 1 and 7"

In Reference 7, when a search result is not obtained, a condition is relaxed in such a manner that the number of characters used as wildcards is increased. However, if the number of characters used as wildcards is increased, a number of other word having only several different characters are detected as search results. That is, a number of other words are likely to be incorrectly retrieved.

In the invention of claim 14, a reference distance in a character element distance table is set to a large value. This means that the number of characters to be subjected to comparison is increased by narrowing a search target to similar characters which tend to be misrecognized. Therefore, it is less possible that extra words are incorrectly retrieved as compared to Reference 7.

Therefore, claim 14 has inventive step over References 1 and 7.

(9) Regarding "claim 15 lacks inventive step over References 1 and 8"

In Reference 8, a search-requested character string is subjected to morphological analysis to be divided into words. A search is performed for the divided words.

**THIS PAGE BLANK (USPTO)**

Therefore, it is possible to search character strings having the same meaning but different expression.

The invention of claim 15 is the same as the subject matter of Reference 8 in the point that a search-requested character string is subjected to morphological analysis to be divided into words. However, in the invention of claim 15, when a search is performed using a probability based on the number of characters (character elements) in a search character string and the number of matched characters (character elements), it is possible to prevent detection of words which are not a search target due to compound words or the like (see Example 14).

Therefore, claim 15 has inventive step over References 1 and 8.

**THIS PAGE BLANK (USPTO)**